UNIVERSITE DU DROIT ET DE LA SANTE - LILLE 2

FACULTE DE MEDECINE HENRI WAREMBOURG

Année: 2016

THESE POUR L'OBTENTION DU GRADE DE DOCTEUR DE L'UNIVERSITE DE LILLE 2

Discipline: Mathématiques

Spécialité: Mathématiques appliquées et applications des mathématiques

Vers la segmentation automatique des organes à risque dans le contexte de la prise en charge des tumeurs cérébrales par l'application des technologies de classification de deep learning

Présentée et soutenue publiquement le 15-Juin à 14h30
au Pole Formation

Par Jose DOLZ

JURY

Président :

Monsieur le Docteur Albert LISBONA

Rapporteurs :

Monsieur le Docteur Pierre JANNIN

Madame la Professeure Su RUAN

Examinateurs :

Monsieur le Docteur Nacim BETROUNI

Monsieur le Docteur Christian DURIEZ

Directeur de Thèse :

Monsieur le Docteur Maximilien VERMANDEL

Membres invités:

Monsieur Laurent MASSOPTIER

Madame la Docteure Hortense A. KIRISLI

Vers la segmentation automatique des organes à risque dans le contexte de la prise en charge des tumeurs cérébrales par lápplication des technologies de classification de deep learning

Résumé : Les tumeurs cérébrales sont une cause majeure de décès et d'invalidité dans le monde, ce qui représente 14,1 millions de nouveaux cas de cancer et 8,2 millions de décès en 2012. La radiothérapie et la radiochirurgie sont parmi l'arsenal de techniques disponibles pour les traiter. Ces deux techniques s'appuient sur une irradiation importante nécessitant une définition précise de la tumeur et des tissus sains environnants. Dans la pratique, cette délinéation est principalement réalisée manuellement par des experts avec éventuellement un faible support informatique d'aide à la segmentation. Il en découle que le processus est fastidieux et particulièrement chronophage avec une variabilité inter ou intra observateur significative. Une part importante du temps médical s'avère donc nécessaire à la segmentation de ces images médicales. L'automatisation du processus doit permettre d'obtenir des ensembles de contours plus rapidement, reproductibles et acceptés par la majorité des oncologues en vue d'améliorer la qualité du traitement. En outre, toute méthode permettant de réduire la part médicale nécessaire à la délinéation contribue à optimiser la prise en charge globale par une utilisation plus rationnelle et efficace des compétences de l'oncologue.

De nos jours, les techniques de segmentation automatique sont rarement utilisées en routine clinique. Le cas échéant, elles s'appuient sur des étapes préalables de recalages d'images. Ces techniques sont basées sur l'exploitation d'informations anatomiques annotées en amont par des experts sur un "patient type". Ces données annotées sont communément appelées "Atlas" et sont déformées afin de se conformer à la morphologie du patient en vue de l'extraction des contours par appariement des zones d'intérêt. La qualité des contours obtenus dépend directement de la qualité de l'algorithme de recalage. Néanmoins, ces techniques de recalage intègrent des modèles de régularisation du champ de déformations dont les paramètres restent complexes à régler et la qualité difficile à évaluer. L'intégration d'outils d'assistance à la délinéation reste donc aujourd'hui un enjeu important pour l'amélioration de la pratique clinique.

L'objectif principal de cette thèse est de fournir aux spécialistes médicaux (radiothérapeute, neurochirurgien, radiologue) des outils automatiques pour segmenter les organes à risque des patients bénéficiant d'une prise en charge de tumeurs cérébrales par radiochirurgie ou radiothérapie.

Pour réaliser cet objectif, les principales contributions de cette thèse sont présentées sur deux axes principaux. Tout d'abord, nous considérons l'utilisation de l'un des derniers sujets d'actualité dans l'intelligence artificielle pour résoudre le problème de la segmentation, à savoir le “deep learning”. Cet ensemble de techniques présente des avantages par rapport aux méthodes d'apprentissage statistiques classiques (Machine Learning en anglais). Le deuxième axe est dédié à l'étude des caractéristiques d'images utilisées pour la segmentation (principalement les textures et informations contextuelles des images IRM). Ces caractéristiques, absentes des méthodes classiques d'apprentissage statistique pour la segmentation des organes à risque, conduisent à des améliorations significatives des performances de segmentation. Nous proposons donc l'inclusion de ces fonctionnalités dans un algorithme de réseau de neurone profond (deep learning en anglais) pour segmenter les organes à risque du cerveau.

Nous démontrons dans ce travail la possibilité d'utiliser un tel système de classification basée sur techniques de “deep learning” pour ce problème particulier. Finalement, la méthodologie développée conduit à des performances accrues tant sur le plan de la précision que de l'efficacité.

Mots clès : Segmentation des organes à risque, radiochirurgie, radiothérapie, réseau de neurones profond.

Towards automatic segmentation of the organs at risk in brain cancer context via a deep learning classification scheme

Brain cancer is a leading cause of death and disability worldwide, accounting for 14.1 million of new cancer cases and 8.2 million deaths only in 2012. Radiotherapy and radiosurgery are among the arsenal of available techniques to treat it. Because both techniques involve the delivery of a very high dose of radiation, tumor as well as surrounding healthy tissues must be precisely delineated. In practice, delineation is manually performed by experts, or with very few machine assistance. Thus, it is a highly time consuming process with significant variation between labels produced by different experts. Radiation oncologists, radiology technologists, and other medical specialists spend, therefore, a substantial portion of their time to medical image segmentation. If by automating this process it is possible to achieve a more repeatable set of contours that can be agreed upon by the majority of oncologists, this would improve the quality of treatment. Additionally, any method that can reduce the time taken to perform this step will increase patient throughput and make more effective use of the skills of the oncologist.

Nowadays, automatic segmentation techniques are rarely employed in clinical routine. In case they are, they typically rely on registration approaches. In these techniques, anatomical information is exploited by means of images already annotated by experts, referred to as atlases, to be deformed and matched on the patient under examination. The quality of the deformed contours directly depends on the quality of the deformation. Nevertheless, registration techniques encompass regularization models of the deformation field, whose parameters are complex to adjust, and its quality is difficult to evaluate. Integration of tools that assist in the segmentation task is therefore highly expected in clinical practice.

The main objective of this thesis is therefore to provide radio-oncology specialists with automatic tools to delineate organs at risk of patients undergoing brain radiotherapy or stereotactic radiosurgery. To achieve this goal, main contributions of this thesis are presented on two major axes. First, we consider the use of one of the latest hot topics in artificial intelligence to tackle the segmentation problem, i.e. deep learning. This set of techniques presents some advantages with respect to classical machine learning methods, which will be exploited throughout this thesis. The second axis is dedicated to the consideration of proposed image features mainly associated with texture and contextual information of MR images. These features, which are not present in classical machine learning based methods to segment brain structures, led to improvements on the segmentation performance. We therefore propose the inclusion of these features into a deep network.

We demonstrate in this work the feasibility of using such deep learning based classification scheme for this particular problem. We show that the proposed method leads to high performance, both in accuracy and efficiency. We also show that automatic segmentations provided by our method lie on the variability of the experts. Results demonstrate that our method does not only outperform a state-of-the-art classifier, but also provides results that would be usable in the radiation treatment planning.

Keywords: Machine learning, support vector machines, deep learning, stacked denoising auto-encoders, radiotherapy,

To my wife Silvia, and my sons Eithan and Noah

“Shoot for the moon, even if you fail, you’ll land among the stars.”

Acknowledgments

I would like to thank to all the people who made my stay in Lille one of the most cherish experiences of my life. Special mention goes to my enthusiastic advisor Dr. Maximilien Vermandel, whose advises and support have been invaluable. My PhD has been an amazing experience and your advices on both research as well as on my current and future career have been priceless. Similar, profound gratitude goes to Laurent Massoptier, who supervised me during all his time in the company I worked in. I am indebted for his faith on my work and also for his support. I would also like to thank my committee members, Dr. Pierre Jannin, Prof. Su Ruan, Dr. Nacim Betrouni, Dr. Albert Lisbona and Prof. Christian Duriez for serving as jury members of my PhD dissertation. Special mention goes to Dr. Pierre Jannin, who have served as president of the PhD monitoring committee since the beginning and who guided to improve this work.

I would especially like to thank experts that participated in this thesis by manually contouring all what we needed. Particularly, I would like to express my gratitude to Prof. Nicolas Reyns, for collecting data for my Ph.D. thesis.

I would like to express my appreciation to all those persons, companies and departments who have offered me their time during the consecution of this research. First, I would like to give my sincere thanks to AQUILAB, specially to David Gibon and Philippe Bourel, for letting me the opportunity to work in their company, where I have grown up both professionally and personally. In addition, there are two people that I need to mention specially, Dr. Romain Viard and Dr. Hortense A. Kirisli. Their friendship and unselfish help enabled me to improve as researcher. I owe them my sincere gratitude for their generous and timely help. Last, I would like thank to all the partners that composed the FP-7 European project SUMMER, particularly to the Department of Radiation Oncology at the University Medical Center in Freiburg, Germany, and the Center for Medical Physics and Biomedical Engineering at the Medical University of Vienna in Vienna, Austria.

Since they are very important on my life, I would like to thank to my parents, my sister and my family in law for their love and support. Specially mention goes to my mother, which support, patience and efforts made me follow the good way.

Finally, I thank will all my love to my wife Silvia and my sons Eithan and Noah. Silvia's support, her encouragement, patience and unwavering love were undeniably the bedrock upon which the past seventeen years of my life have been built. I cannot imagine a more special soul-mate; I cannot imagine a better mother for my children. They give the strength I need every day and to whom this dissertation is dedicated.

The work presented in this dissertation has received funding from the European Union s Seventh Framework Programme for research, technological development and demonstration under grant agreement no PITN-GA-2011-290148.

Contents

Chapter 1

Overview

1.1 Context

Cancer is a leading cause of death and disability worldwide, accounting for 14.1 million of new cancer cases and 8.2 million deaths in 2012 [1]. Cancer represents a group of common, non-communicable, chronic and potentially lethal diseases affecting most families in developed countries, and a growing contributor to premature death within population of these countries [2]. Meanwhile, the annual incidence of cancer keeps raising with an estimation of 26 million of new cases yearly by 2030, with a death toll close to 17 million people [3]. In particular, brain tumors are the second most common cause of cancer death in men ages 20 to 39 and the fifth most common cause of cancer among women age 20 to 39 [4].

Among available techniques to treat brain tumors, radiotherapy and radio surgery have become part of the management of patients with brain tumors, to complement surgery or chemotherapy. During treatment, high intensity radiation beams to destroy the cancerous cells are delivered across the tissues. However, when delivering radiation through the human body, side effects on normal tissues may occur. To limit the risk of severe toxicity of critical brain structures, i.e. the organs at risk (OARs), the volume measurements and the localization of these structures are required. Among available image modalities, magnetic resonance imaging (MRI) images are extensively used to segment most of the OARs, which is performed mostly manually nowadays. However, manual delineation of brain structures is prohibitively time-consuming, and might never be reproducible during clinical routines [5,6], leading to substantial inconsistency in the segmentation.

Medical imaging is increasingly evolving towards higher resolution and throughput. The exponential growth of the amount of data in medical imaging and the usage of multiple imaging modalities have significantly increased the need of computer assisted tools in clinical practice. Among them, automatic segmentation of brain structures has become a very important field of the medical image processing research. A variety of techniques has been presented during the last decade to segment brain structures. Particularly, structures involved in neurological diseases, such as Alzheimer or Parkinson, have held the attention of researchers. However, critical structures involved in the radiation treatment planning are rarely included in the evaluations. Even in the cases they are analyzed, limited success has been reported. Nevertheless, the fields of computer vision and machine learning are closely related to offer a rich set of useful techniques in the medical imaging domain, in general, and in segmentation in particular.

In this thesis, deep learning techniques are proposed as alternative to the segmentation of the OARs to address the problems of classical segmentation methods. Specifically, an unsupervised deep learning technique known as Stacked Denoising Auto-Encoders is proposed and evaluated. The application of SDAE to the segmentation of OARs in brain cancer allows to (*i*) yield more accurate classification in more complex environments, (*ii*) achieve faster classification without sacrifying classification accuracy and (*iii*) avoid expensive registration stages.

1.2 Contributions

The main contributions of this thesis can be summarized as follows:

- An unsupervised deep learning technique known as Stacked Denoising Auto-Encoders is proposed to segment the OARs in radiotherapy and radio-surgery as alternative to conventional methods used to segment brain structures, i.e. atlas-based.

- New features to include in the classification scheme are proposed to improve the performance of other researchers that used traditional features in Machine Learning classification schemes. Some of the proposed features, have been already employed in neuroimaging. However, their use is limited to other applications rather than segmentation of the OARs.

- Some OARs that have not been previously segmented by proposed methods are included in the list of OARs involved.

- The proposed deep learning classification scheme is compared to a well-known state-of-the-art machine learning classifier, support vector machines.

- Apart from the previous technical validation of the presented approach, its performance is evaluated in clinical routine. Four observers contributed in this thesis by doing manual contouring of all the OARs involved in the radiation treatment planning (RTP). Results provided by the automatic method were compared to the manual ones.

1.3 Roadmap

This dissertation is organized as follows.

In Chapter 2, an introduction of brain cancer, radiation techniques and effects of radiation on biological tissues will be presented. As a part of the radiation treatment planning, the problem of manual segmentation of the organs at risk and the blueneed of automatizing this step will be introduced.

In Chapter 3, the relevant literature on the state-of-the-art segmentation methods for brain structures on MRI is presented. Particularly the development of atlas-based methods, statistical models of shape and texture, deformable models and machine learning techniques are reviewed. To reduce the number of meaningless papers, only works handling with brain structures that are included in the RTP are considered.

In Chapter 4, the main contributions of this work are disclosed. First section brings to the reader a theoretical introduction of machine learning basis terms. Concepts such as classification or data representation are briefly explained. Next, historical context, advantages and explanation of deep learning, and particularly the technique employed in this thesis, are afforded. Afterwards, the proposed features to segment brain structures in this dissertation are detailed. Last two sections of this chapter presents the methodological processes performed in this thesis to conduct both training and classification.

In Chapter 5, we detail the materials employed throughout this thesis. Image modalities and their characteristics, as well as volume contours used as reference are introduced. Afterwards, strategies and metrics used to evaluate the performance of the proposed classification system are presented.

In Chapter 6, experiments set-up and results of these experiments are shown. Comparisons with other works is also conducted in this chapter.

In chapter 7, conclusions of the methods presented in this thesis, as well as guidelines for future work are discussed.

And finally, chapters 8 and 9 present the scientific dissemination produced by this work, and a french summary of the thesis, respectively.

Chapter 2

Introduction

"What we do in life, echoes in eternity."

2.1 Brain Cancer

Cancer is a leading cause of death and disability worldwide, accounting for 14.1 million of new cancer cases and 8.2 million deaths in 2012 [1]. Cancer represents a group of common, non-communicable, chronic and potentially lethal diseases affecting most families in developed countries, and a growing contributor to premature death within population of these countries [2]. Meanwhile, the annual incidence of cancer keeps raising with an estimation of 26 million of new cases yearly by 2030, with a death toll close to 17 million people [3]. In particular, brain tumors are the second most common cause of cancer death in men ages 20 to 39 and the fifth most common cause of cancer among women age 20 to 39 [4].

A brain tumor is any mass caused by abnormal or uncontrolled growth of cells that arise within or adjacent to the brain. In general, these tumors are categorized according to several factors, including location, type of cells involved, and the growing rate. Slowly growing tumors that lack of capacity to spread to distant sites and that originate in the brain itself are called primary brain tumors. On the other hand, rapidly growing tumors that can infiltrate surrounding tissues and spread to distant sites, i.e. metastasize, are called secondary brain tumors. While primary brain tumors can be benign or malignant, secondary brain tumors are always malignant. However, both types are potentially disabling and life threatening. Because the space inside the skull is limited, their growth increases intracranial pressure, and may cause edema, reduced blood flow, and displacement, with consequent degeneration of healthy tissue that controls vital functions [7, 8]. Additionally, metastatic or secondary brain tumors are the most common types of brain tumors, and occur in 10-15 % of people with cancer. Brain tumors are inherently difficult to treat given that the unique features of the brain can complicate the use of conventional diagnostic and treatment methods.

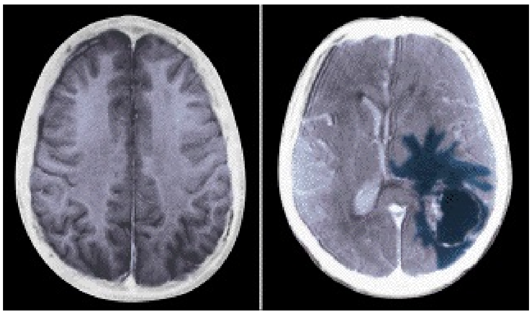


Figure 2.1: Healthy brain(left) compared to brain tumor (in blue,right).

2.2 Brain Tumor Treatment

One of the consequences of tumor growing into or pressing on a specific region of the brain is the probability of stopping that brain area from working the way it should. Consequently, independently on the nature of the tumor, both benign and malignant brain tumors cause signs and symptoms and require treatment.

2.2.1 Available Treatments

A variety of therapies are used to treat brain tumors. Treatments options mainly include surgery, radiotherapy, chemotherapy, and/or steroids. Selection of suitable treatments depends on a number of factors, which may include type, location, size or grade of the tumor, as well as the patient′s age and general health. Surgery is used to excise tumors, or parts of tumors, from specific locations directly using a knife. Chemotherapy uses chemical substances to treat cancer indirectly, since these drugs typically target all rapidly dividing cells, which include cancer cells. Radiation therapy (RT) uses radiation to kill tumor cells, which involves radiation permanently damaging the deoxyribonucleic acid (DNA) of tumor cells.

2.2.2 Radiation Therapy

The term radiation therapy, or radiotherapy (RT), describes the medical application of ionizing radiation to control malignant cells by damaging their DNA [9]. Essential genetic instructions for the development and functioning of a cell are contained in the DNA. Cells are naturally programmed to correct damaged DNA up to a certain degree. Nevertheless, if the deterioration is substantial, the cell dies. It has been demonstrated, however, that healthy cells recover better than cancerous cells when they are exposed to degradation [10]. This radiobiological difference between healthy and cancerous cells is exploited by radiation therapy. An example of a brain tumor patient having been treated with RT is shown in figure 2.2.

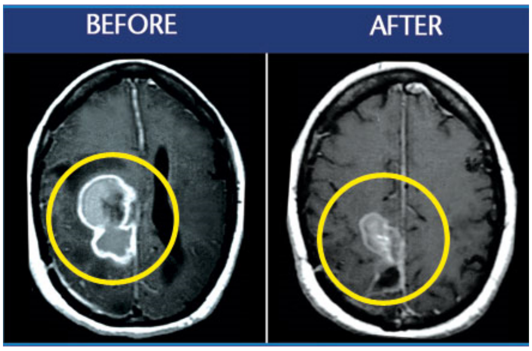


Figure 2.2: A patient before and after of having being treated with rdiation therapy.

The three primary techniques for delivering radiation include: i) external or conventional radiotherapy, ii) internal radiotherapy or brachytherapy, and iii) stereotactic radiosurgery (SRS), sometimes referred to as gamma-knife. Each of them have been evaluated in the treatment of patients with brain tumors and may be utilized in different circumstances. While external radiotherapy is the conventional treatment for brain tumors, SRS has also become a standard procedure. Recently, SRS has been used in the treatment of many types of brain tumors, such as acoustic neuromas, meningiomas or trigeminal neuralgia, for example. Furthermore, it has been proven to be effective in the treatment of brain metastases. Since this work aims at improving the segmentation procedure in RT and SRS treatment planning, only these two techniques will be explained in the following section.

2.2.2.1 Conventional Radiotherapy

RT involves directing radiation beams from outside the body into the tumor. It implicates careful and accurate use of high intensity radiation beams to destroy the cancerous cells. Machines called linear accelerators (LINAC) produce these high energy radiation beams which penetrate the tissues and deliver the radiation dose deep in the body where the tumor is located. These modern machines and other state-of-the-art techniques have enabled radiation oncologists to enhance the ability to deliver radiation directly to the tumor whilst substantially reducing the side effects.

RT is typically delivered as an outpatient procedure for approximately over a six to eight week period, five days a week. Nevertheless, treatment schedule may vary across patients. The total procedure for each session typically takes

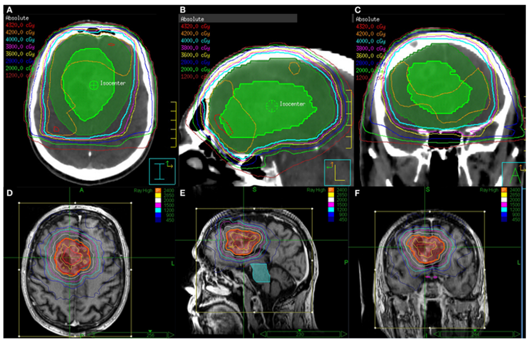


Figure 2.3: Conventional RT and CyberKnife SRS treatment plans for a patient who received 40 Gy in 15 fractions to FLAIR for the first course followed an SRS boost to T1 Enhancement at a total dose of 24 Gy delivered in 3 fractions. Shown are the (A) axial, (B) sagittal, and (C) coronal views of the EBRT treatment plans and the (D) axial, (E) sagittal, and (F) coronal views of the CyberKnife SRS treatment plans.

between 10 and 20 minutes. This dose fractionation enables normal tissue to recover between two fractions reducing damage to normal tissues. RT begins with a planning session during which the radiation oncologist places marks on the body of the patient and takes measurements in order to align the radiation beam in the precise position for each treatment. During treatment, the patient lies on a table and the radiation is delivered from multiple directions to minimize the dose received by healthy tissues. A conventional RT and CyberKnife SRS treatment plan are shown in Figure 2.3 (Image courtesy of [11]).

2.2.2.2 Stereotactic Radiosurgery

Stereotactic techniques have been developed with the aim to deliver more localized irradiation and minimize the long-term consequences of treatment. They represent a refinement of conventional RT with further improvement in immobilization, imaging and treatment delivery. Basically, SRS is a single fraction RT procedure at high dose. For instance, while a dose of 2 Gy is delivered for a standard RT fraction, 12 to 90 Gy are delivered in a SRS fraction. Thus, the entire procedure occurs in one day, including immobilization, scanning, planning and the procedure itself.

When a patient undergoes SRS, the radiation dose delivered in one session is commonly lower than the total dose that would be given by following conventional RT. Nevertheless, the tumor receives a very high radiation does at once with SRS. Since more radiation is delivered to surrounding healthy tissues when treatment is split into few or several sessions instead of one, decreasing the number of sessions is important. Otherwise, it might result in more side effects, some of which may be permanent. Other consequence of splitting the treatment is that, a reduced amount of radiation delivered to the tumor with each RT session, rather than a very large dose in a single session, may result in less tumor control and poorer outcomes than by employing SRS.

Even though RT and SRS are reported to have identical outcomes for particular indications [12] and regardless of similarities between their concepts, the intent of both approaches is fundamentally different. On the one hand, conventional RT relies on a different sensitivity of the target and the surrounding normal tissue to the total accumulated radiation dose [13]. On the other hand, SRS aims at destroying target tissue while preserving adjacent normal tissue. In other words, SRS offers the possibility of normal tissue protection by improved precision of beam application, while conventional RT is limited to the maximum dose that can be safely applied because of normal tissue constraints. Instead of many doses of radiation therapy to treat a targeted region, SRS usually consists of a single treatment of a very high dose of radiation in a very focused location. Due to this, not only higher total radiation doses but also higher single doses can be used, which results in increased biologically effective doses compared with conventional RT.

Stereotactic radiosurgery is a well-described management option for most metastases, meningiomas, schwannomas, pituitary adenomas, arteriovenous malformations, and trigeminal neuralgia, among others [12, 14].

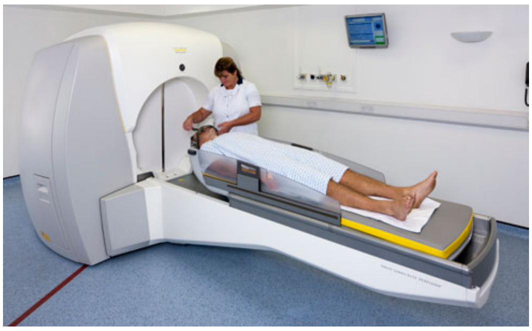


Figure 2.4: A patient being positioned for SRS treatment (Gamma-Knife).

The popularity and acceptance of SRS procedures has led to the development of several SRS systems. Stereotactic boosts can be carried out in several modalities, such as Gamma Knife (Elekta AB, Stockholm, Sweden), and various LINAC-based systems such as CyberKnife (Accuray Inc., Sunnyvale, CA) or Novalis (BrainLAB, Feldkirchen, Germany).

2.2.2.2.1 Gamma Knife. The Gamma Knife (GK) is an instrument that was developed by surgeons in Sweden nearly five decades ago. A GK typically contains 201 beams of highly-focused gamma rays that are directed so that they intersect at the precise location of the cancer. The patient is placed on a couch and then a specialized helmet (Fig. 2.5) is attached to the head frame. Holes in the helmet allow the beams to match the calculated shape of the tumor.

The most frequent use of the Gamma Knife has been for small, benign tumors, particularly acoustic neuromas, meningiomas, and pituitary tumors. In addition, the GK is also employed to treat solitary metastases and small malignant tumors with well-defined borders.

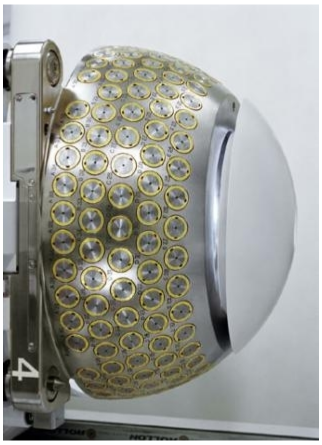


Figure 2.5: Gamma Knife radiation helmet.

2.2.2.2.2 Linear accelerators (LINAC). Although a linear accelerator (LINAC) is mostly employed for conventional RT treatments, some SRS system have adopted its use to treat brain cancer patients. A LINAC customizes high energy x-ray beams to conform to a defined tumor's shape. The high energy x-rays are delivered to the region where the tumor is present. The patient is positioned on a sliding bed around which the linear accelerator circles. The linear accelerator directs arcs of x-ray beams at the tumor. The pattern of the arc is computer-matched to the tumor's shape. This reduces the dose delivered to surrounding normal tissue. The LINAC can perform SRS on larger tumors either during multiple sessions, which is referred to as fractionated stereotactic radiotherapy.

2.2.3 Radiation Treatment Flowchart

Radiation treatment planning (RTP) is often organized in two phases: the planning and the delivery. Images are first acquired, the regions of interest are identified and the ballistic problem is solved for the acquired data. The planned treatment is then delivered to the patient.

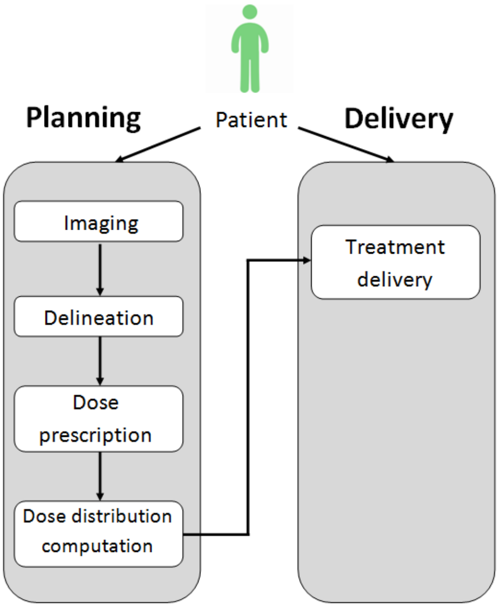


Figure 2.6: Flowchart of a common radiotherapy treatment.

2.2.3.1 Imaging

The CT image gives an estimation of the electronic density of the anatomy, which is still required to compute the dose distribution in the patient body. Since this image modality is affected by a lack of contrast between soft tissues, other images have sometimes to be acquired. Depending on the cancer type, other images such as positron emission tomography (PET) or magnetic resonance imaging (MRI) can be recommended. A detailed justification of the importance of MRI in brain cancer is explained in Section 2.4.

2.2.3.2 Delineation

Acquired images are used to determine the position of the target volumes (TVs) as well as the position of some specific organs. This task is usually performed by the physician. To determine the position of the TVs, the physician defines the borders of regions of interest on the image that corresponds to the gross tumor volumes (GTVs). This operation is known as delineation. It is generally performed by drawing contours on two dimensional (2D) slices extracted from the 3D CT. The delineated region of interest, is made up of several 2D shapes from different slices of the image. As there are assumptions of microscopic spread of cancerous cells around the tumors, margins are added around the GTV. The new volume, called clinical target volume (CTV), takes into account cancerous cells that may not be seen on the image. A third volume, the planning target volume (PTV), is created as an extension of the CTV and takes into account the uncertainties in planning and treatment delivery. It is a geometric volume designed to ensure that the prescribed radiation dose is correctly delivered to the CTV. Critical organs have to be delineated to ensure that they do not receive a higher-than-safe dose. There exist different specifications for each of the organs. In some cases, as for the PTV, an extra margin is added around the organ to take into account the uncertainties. Depending on the localization of the tumor, the delineation stage can take up to 2 hours.

2.2.3.3 Dose prescription

During this stage the physician evaluates the tumor propagation in the patient body by using staging system such as ”tumor-nodes-metastasis” (TNM) and makes the appropriate prescription. The prescription includes, among others, the number of fractions and the dose the tumour has to receive. Those prescriptions must follow the recommendations made by the International Commission on Radiation Units and Measurements (ICRU) (reports ICRU 50, ICRU 62 and ICRU 83).

2.2.3.4 Dose distribution computation

The delineated images and the prescriptions are then given to the physicist who computes the dose distribution. The physicist tries to find the best trade-off between maximizing the dose on the PTV and preserving the critical healthy structures.

2.2.3.5 Treatment Delivery

According to the treatment modality selected, treatment delivery will be either fractionated during several weeks, with one daily session without including the weekend, or delivered in a single session. Regardless of the treatment technique used, during each of these sessions, the patient receives a fraction of the planned dose.

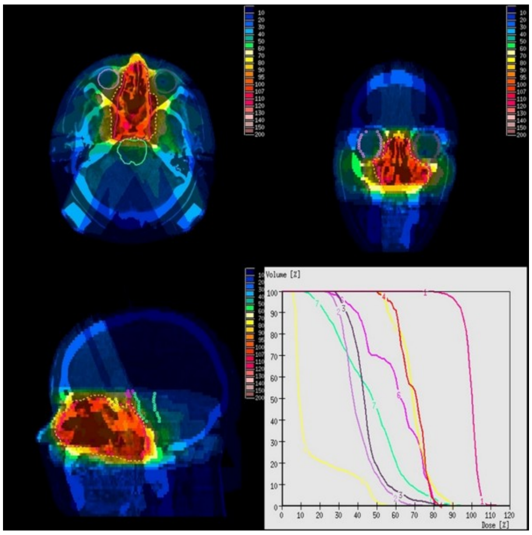


Figure 2.7: Transversal, coronal and sagittal dose distribution and DVH information. Graphs: PTV (1), left eye (2), right eye (3), right optic nerve (4), left optic nerve (5), chiasma (6), brainstem (7), spinal cord (8).

2.3 Effects of radiation on biological tissues

A major goal of RT is to deprive cancer cells of their multiplication potential and eventually kill the cancer cells. However, radiation will also damage healthy cells. Hence, the main goal of a radiation therapy treatment becomes to deliver sufficient dose to the tumor, while ensuring that the healthy tissue around the tumor is spared as much as possible. Particularly in treatments that include SRS, where radiation dose is considerably higher, setup or localization errors might result in severe overdosing of the adjacent normal healthy tissue. This over exposition to radiation may lead to progressive and irreversible complications to the brain, which often occur months or years after treatment. These critical structures to be preserved are referred to as Organs at Risk (OARs).

To deliver the correct radiation dose, the radiation oncologist or neurosurgeon must consider not only the effects of treatment on the tumor but also the consequences on normal tissues. These two objectives cannot be fully achieved simultaneously, because both the probability of undesirable effects of radiotherapy on normal tissues and the probability of tumor control increase with the delivered dose (Figure 2.8). The two sigmoid curves respectively refer to the tumor control probability (TCP, grey curve) and to the normal tissue complication probability (NTCP, red curve). In clinical applications, the effectiveness of radiotherapy is measured by the therapeutic ratio (TCP/NTCP) which ideally should be as high as possible. Typical values in a good radiotherapy treatment are higher than 0.5 for the TCP, and lower than 0.05 for the NTCP.

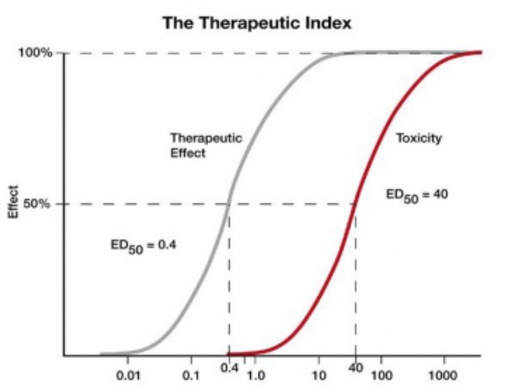


Figure 2.8: The principle of therapeutic ratio. Grey curve represents the TCP, and red curve the probability of complications. The total clinical dose is usually delivered in 2Gy fractions in EBRT.

2.3.1 Organs at Risk

During radiotherapy treatment planning, the normal tissues / critical organs within the radiation beam and at the vicinity of the tumor receive a higher amount of radiation dose, and sometimes may be equal to the tumor dose, which causes normal tissue injury.

The focus of this section is therefore on providing a background in the anatomy that underlies the images that we are attempting to segment. Understanding the role of each of these organs is crucial to comprehend how an overdose may damage their primary functions leading to a decrease of the life's quality of the patient.

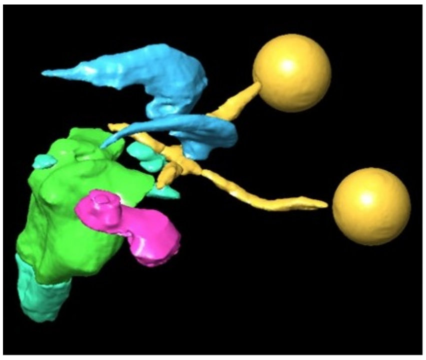


Figure 2.9: Organs at Risk commonly involved in brain tumor radiation treatment.

2.3.1.1 Brainstem

The brainstem, or brain stem, is one of the most basic regions of the human brain. Despite this, it is one of the most vital regions for our body's survival. It represents one of the three major parts of the brain, which controls many important body functions. In the anatomy of humans it is the posterior part of the brain, adjoining and structurally continuous with the spinal cord. It is usually described as including the medulla oblongata (myelencephalon), pons (part of metencephalon), and midbrain (mesencephalon). Though small, this is an extremely important part of the brain as the nerve connections of the motor and sensory systems from the main part of the brain to the rest of the body pass through the brainstem. This includes the corticospinal tract (motor), the posterior column-medial lemniscus pathway (fine touch, vibration sensation, and proprioception), and the spinothalamic tract (pain, temperature, itch, and crude touch). The brainstem also plays an important role in the regulation of cardiac and respiratory function. It also regulates the central nervous system, and is pivotal in maintaining consciousness and regulating the sleep cycle.

2.3.1.2 Eyes

Eyes are the organs of vision. They detect light and convert it into electrochemical impulses in neurons. The different parts of the eye allow the body to take in light and perceive objects around us in the proper color, detail and depth. This allows people to make more informed decisions about their environment. If a portion of the eye becomes damaged, one may not be able to see effectively, or lose vision all together.

Optic nerves join about half way between the eye and brain, and then split up again. The join is called the optic chiasm. At the join, signals from the 'nose' side of each eye's visual world swap sides and continue traveling along the opposite side from where they started. The two optic nerves then join on to the brain. The brain is split into two halves, right and left. This means all the signals from the visual world on the right hand side are now traveling in the left side of the brain. It also means that all the signals from the visual world on the left hand side are now traveling in the right half of the brain.

The information then travels to the many different special 'vision' areas of the brain. The main bit of the brain that works vision is at the back of the head. It is called the occipital lobe. The joined up path that signals travel down from retina to optic nerve then optic chiasm then occipital lobe is called the visual pathway. There are two visual pathways, one on the right side of the brain and another on the left. All parts of both visual pathways need to be present and working for us to see normally.

2.3.1.3 Optic Nerves

The optic nerves are located in the back of the eyes. However, although the optic nerve is part of the eye, it is considered to be in the central nervous system. The optic nerve is the nerve that carries the neural impulses created by the retina to the brain, where this information is interpreted. At a structure in the brain called the optic chiasm, each optic nerve splits, and half of its fibers cross over to the other side. The crossing over of optic nerve fibers at the optic chiasm allows the visual cortex to receive the same hemispheric visual field from both eyes. Superimposing and processing these monocular visual signals allow the visual cortex to generate binocular and stereoscopic vision.

Any damage or disorder on the optic nerves will always impact vision in some way and might affect either one or both eyes.

2.3.1.4 Optic Chiasm

The optic chiasm is located in the forebrain directly in front of the hypothalamus. Crucial to sight, left and right optic nerves intersect at the chiasm. One-half of each nerve's axons enter the opposite tract at this location, making it a partial decussation.

We have seen that the optic nerves send electrical signals from each eye to meet in the brain at the optic chiasma. Here, the left visual signal from one eye is combined with the other eye and the same goes for the right visual signal. Now the signals split again. The right visual heads for the left brain and the left visual makes its way to the right side of the brain. This way, visual messages from both eyes will reach both halves of the visual cortex. The brain then merges the image into one image which you are looking out at the world with. This partial crossing of the nerve fibers at the optic chiasm (or chiasma) is the reason why we humans have stereoscopic sight and a sense of depth perception.

2.3.1.5 Pituitary Gland

The pituitary gland is a pea-sized structure located at the base of the brain, just below the hypothalamus and attached to it by nerve fibers. It is part of the endocrine system and produces hormones which control other glands as well as various bodily functions. The pituitary is divided into three sections known as the anterior, intermediate and posterior lobes, each of which produces specific hormones. The anterior lobe is mainly involved in development of the body, sexual maturation and reproduction. Hormones produced by the anterior lobe regulate growth and stimulate the adrenal and thyroid glands as well as the ovaries and testes. It also generates prolactin, which enables new mothers to produce milk. The intermediate lobe of the pituitary gland releases a hormone which stimulates the melanocytes, cells which control pigmentation through the production of melanin. The posterior lobe produces antidiuretic hormone, which reclaims water from the kidneys and conserves it in the bloodstream to prevent dehydration. Oxytocin is also produced by the posterior lobe, aiding in uterine contraction during childbirth and stimulating the production of milk.

2.3.1.6 Hippocampus

The hippocampus is a small region of the brain that belongs to the limbic system and is primarily associated with memory and spatial navigation. The hippocampus is located in the brain's medial temporal lobe, underneath the cortical surface. Its structure is divided into two halves which lie in the left and right sides of the brain. The hippocampus is responsible for long-term, or "declarative" memory, and spatial navigation. Long term memory is like a compilation of data in our conscious memory and all of our gathered knowledge and experiences. The hippocampus is involved in the storage of all of this data. In some neurological disorders, such as Alzheimer's disease, the hippocampus is one of the first regions of the brain to become damaged and this leads to the memory loss and disorientation associated with the condition. Individuals with hippocampal damage develop amnesia and may be unable to form new memories of the time or location of an event, for instance.

2.3.2 Dose limits

For the OARs typically involved in RTP some of the tolerance limits are presented in table 2.1.

2.4 The role of Structural MRI in brain tumor radiation treatment

During the last decades, medical imaging, which was initially used for basic visualization and inspection of anatomical structures, has evolved to become an essential tool for diagnosis, treatment and follow-up of patient diseases. Particularly, in oncology, image evolution has improved the understanding of the complexities of cancer biology, cancer diagnosis, staging, and prognosis. Advanced medical imaging techniques are thus used for tumor resection surgery (i.e. pre-operative planning, intra-operative, post-operative), and for

	Dose level limit(D$_{max}$)
OAR	**Radiotherapy**
Hippocampus	16Gy (IMRT - fractionation 10x3Gy) [15]
Brainstem	45Gy (IMRT - fractionation 20x1.8Gy + 10x(1.8Gy+1.6Gy)) [16]
Eyes(Retina)	40Gy (IMRT - fractionation 30x2Gy) [17]
Eyes(Lens)	As low as possible [17]
Cochlea	45Gy (conventionally fractionated RT) [18]
Chiasma	54Gy (IMRT - fractionation 30x2Gy) [17]
Optic Nerve	54Gy (IMRT - fractionation 30x2Gy) [17]
OAR	**Radiosurgery**
Hippocampus	-
Brainstem	volume 0.1 cc / Dose limit = 10Gy [19]
	volume 0.1 cc / Dose limit=12Gy [20]
Eyes(Retina)	5Gy [21]
Eyes(Lens)	3Gy [21]
Cochlea	12Gy [19]
	10Gy [22]
Chiasma	volume 0.2CC / Dose limit = 8Gy [19]
Optic Nerve	volume 0.2CC / Dose limit = 8Gy [19, 23–25]

Table 2.1: Dose limits for the OARs in both radiotherapy and radio-surgery.

subsequent radiotherapy treatment planning (RTP). There exists a wide range of medical imaging modalities that allows neuro-scientists to see inside a living human brain. Early imaging methods, invasive and sometimes dangerous, have been abandoned in recent times in favor of non-invasive, high-resolution modalities, such as computed tomography (CT), and especially structural magnetic resonance imaging (MRI). However, to outline the normal brain structures in great detail, the MRI has a higher sensitivity for detecting the presence of, or changes within, a tumor. It is therefore perfectly suited for anatomic visualization of the human body such as deep structures and tissues of the brain. For this reason, and because MRI does not rely on ionizing radiation, MRI has gradually supplanted CT as the mainstay of clinical neuro-oncology imaging, becoming the preferred modality for the diagnostic, follow-up and planning treatments of brain lesions [26].

Additional advantage of MRI is offered by the ability to directly obtain images in planes other than axially, as with CT. The high contrast resolution noted with MRI over CT offers better clarity and easier diagnosis and demarcation of soft tissues or lesions in most situations. We can therefore say that structural Magnetic Resonance Imaging plays a central and crucial role in brain tumor radiation treatment (RT) assessment.

The typical MR scan for a patient with a brain tumor includes T1/T2-weighted, fluid-attenuated inversion recovery (FLAIR), and post-contrast T1-weighted images (Figure 2.10). T1-weighted images are most useful for depicting anatomic detail and show cerebrospinal fluid and most tumors as low signal intensity, whereas areas of fat and subacute hemorrhage appear as high signal intensity. T2-weighted images are more sensitive for lesion detection and show cerebrospinal fluid and most lesions as high signal intensity, whereas areas of hemorrhage or chronic hemosiderin deposits may appear as low signal. FLAIR images are T2-weighted with low signal cerebrospinal fluid, are highly sensitive for pathology detection, and display most lesions, including tumors and edema, with higher signal intensity than T2 images. However, the tumor focus in FLAIR or T2 images is not well separated from surrounding edema, gliosis, or ischemic changes. T1-weighted images after contrast enhancement generally provide better localization of the tumor nidus and improved diagnostic information relating to tumor grade, blood-brain barrier breakdown, hemorrhage, edema, and necrosis. Contrast-enhanced T1-weighted images also show small focal lesions better, such as metastases, tumor recurrence, and ependymal or leptomeningeal tumor spread. The T1-weighted enhancement of a contrast agent is attributed to blood-brain barrier leakage associated with angiogenesis and capillary damage in regions of active tumor growth and radiation injury [27].

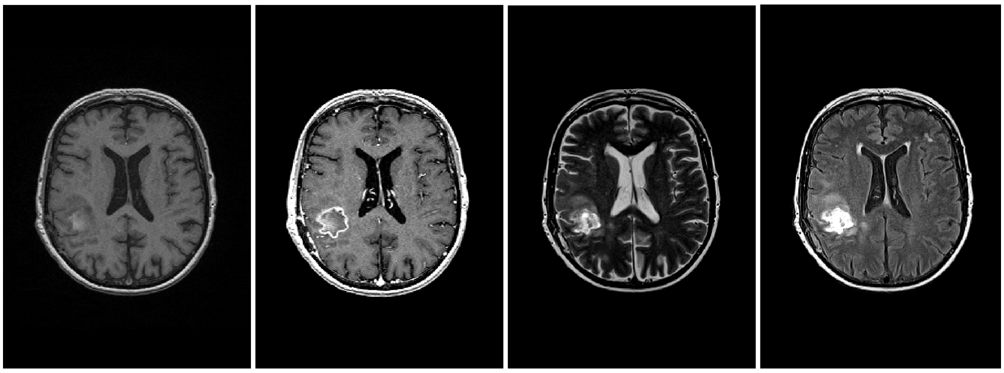


Figure 2.10: MRI modalities commonly employed in the RTP. From left to right: T1, T1-Gadolinium, T2 and FLAIR modalities.

MRI imaging sequences are composed of multiple slices, which positions and thickness might be different from one modality to another, as shown in Figure 2.11. The red, blue and green rectangles refer to commonly used imaging directions to the MRI slices.

The fact that most cranial contouring is performed on the MRI means that an excellent registration between the CT and MRI scans is essential in order to have confidence in the position of the contours during dose calculation. In general the skull provides a good reference point which prevents too much deformation of the cranium, allowing good results to be achieved using rigid registration techniques. However, because of the long acquisition times of MRI scans, the patient couch is typically designed with greater comfort in mind than the RT treatment couch, and this can mean there is some deformation

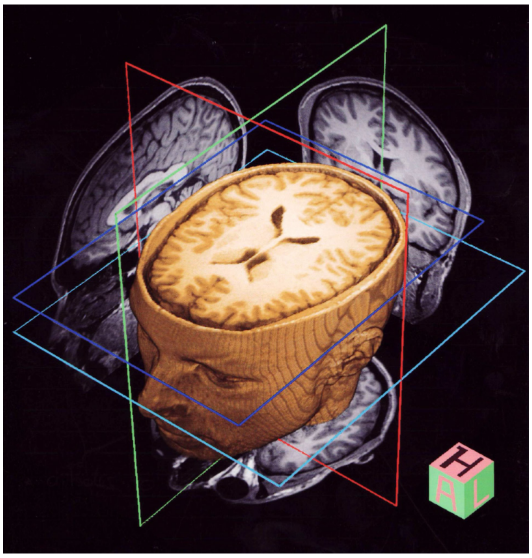


Figure 2.11: The selection of directions of MRI slices.

in the neck area, which can make an overall good fit hard to achieve, instead the oncologist must choose which region to prioritize in the fitting.

2.5 Need of automatization of the OARs segmentation process

Because RT and SRS involve the delivery of a very high dose of radiation, both tumor and surrounding tissue must be precisely delineated. Particularly for the OARs, their volume measurements and localizations are required to constrain the risk of severe toxicity. These segmentations are therefore crucial inputs for the RTP, in order to compute the parameters for the accelerators, and to verify the dose constraints.

As it has been previously discussed, among available image modalities MRI images are extensively used to segment most of the OARs. The delineation task performed manually by experts, or with very few machine assistance [28], is highly time consuming, and there exists significant variation between the labels produced by different experts [29, 30]. For some OARs with clearly defined boundaries these are likely to be on the order of only a few voxels, but for many organs with reduced contrast a difference of 2 cm or more between contour lines is not uncommon, creating large variations in the volumes contoured by different oncologists. Radiation oncologists, radiology technologists, and other medical specialists spend, therefore, a substantial portion of their time to medical image segmentation. Furthermore, recent investigations have shown that the effects of inter-variability in delineating OARs have a significant dosimetric impact [31].

Consequently, the role of delineating contours on a patient's MRI scan is a highly skilled one, which must be carefully supervised by the physician in charge of treatment. The mean time typically spent to analyze and delineate OAR on a brain MRI dataset has been evaluated to 86 min [5], engaging valuable human resources.

If by automatizing this process it is possible to achieve a more repeatable set of contours that can be agreed upon by the majority of oncologists this would improve the quality of treatment. Additionally, any method that can reduce the time taken to perform this step will increase patient throughput and make more effective use of the skills of the oncologist.

Uncertainty and Choice. Contours approved by the oncologist or requiring minor tricks for a few features is highly expected for any automatic segmentation process. indeed, if the physician spends more time making modifications than it would have taken them to contour by hand, then the purpose of the segmentation algorithm is lost.

To overcome these major issues, various computer-aided systems to (semi-) automatically segment anatomical structures in medical images have been developed and published in recent years. However, brain structures (semi-) automatic segmentation still remains challenging, with no general and unique solution. Because all the aforementioned reasons, and as the number of patients to be treated increases, OARs cannot always be accurately segmented, which may lead to suboptimal plans [32]. This makes the introduction in clinical routine of an automated OARs segmentation assisted tool highly desirable.

Chapter 3

Segmentation methods for brain structures: State of the art

" The most difficult thing is the decision to act, the rest is merely tenacity."
Amelia Earhart

This chapter provides an overview of the state of the art in the field of segmentation of brain structures. Methods referenced in this chapter are applied in various fields, not being restricted to radiotherapy. However, despite the large number or techniques proposed to segment different regions of the brain, only those approaches focusing on the critical structures detailed in 2.3.1 are included.

3.1 Introduction

Image segmentation represents the problem of partitioning an image in a semantically purposeful way. Subdivision of the image into meaningful regions allows that a compact and easier representation of the image can be achieved. Grouping of the pixels is done according to a predefined criterion. This criterion can be based on many factors, such as intensity, color, or texture similarities, pixel continuity, and some other higher level knowledge about the objects model. For many applications, segmentation reduces to find an object in a given image. This involves partitioning the image only into two classes of regions. These two classes can be either the object or the background (Fig. 3.1). Thus, image segmentation is often an essential step in further image analysis, object representation, visualization and many other image processing tasks.

3.2 Medical imaging segmentation

Since image segmentation plays a central role in retrieving meaningful information from images, the effective extraction of all the information and features contained in multidimensional images is of increasingly importance in this field. Medical field provides an interesting source of images. In their raw

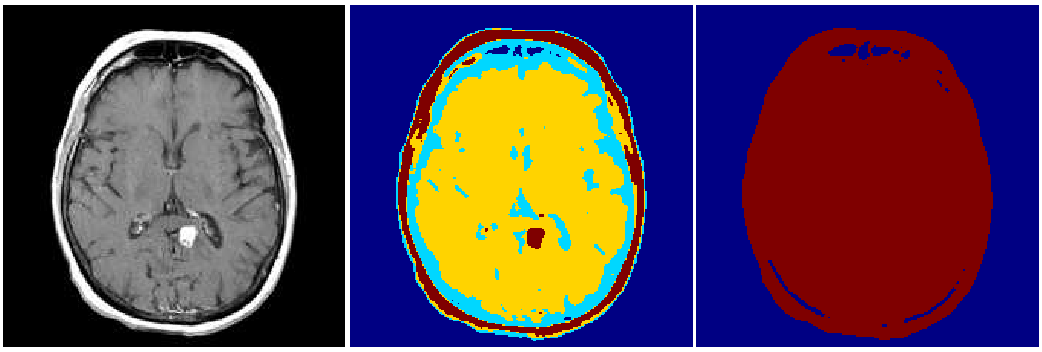


Figure 3.1: Image segmentation example. Original image (*left*) is segmented into four class regions (*center*), and into two class regions (*right*).

form, medical images are represented by arrays of numbers depicting quantities that show contrast between different types of body tissue. Voxel values may vary depending on the image modality, type of tissue or some acquisition parameters. Processing and analyzing medical images are useful to transform this raw information into a quantifiable symbolic form. The extraction of this meaningful quantitative information can aid in diagnosis, as well as in integrating complementary data from multiple imaging modalities. Therefore, in medical image analysis, segmentation has a great clinical value since it is often the first step in quantitative image analysis. For instance, segmentation of medical images aims at identifying the target anatomy or pathology and delineating the boundary of structures of interest for computer aided diagnosis (CAD) purpose or for planning therapy. Image segmentation plays, therefore, an important role in numerous medical applications [33].

However, medical image segmentation distinguishes itself from conventional image segmentation tasks and still remains generally challenging. First, many medical imaging modalities generate very noisy and blurred images due to their intrinsic imaging mechanisms. Particularly, in radiation oncology, radiologists tend to reduce acquisition times on CT and MRI for better patient acceptance. Second, medical images may be relatively poorly sampled. Many voxels may contain more than only one tissue type, which is known as Partial Volume Effect (PVE) (See figure 3.2). When this occurs, the intensity of a given voxel depends not only on the tissue properties, but also on the proportions of each tissue type present in the voxel. As a consequence, loss of contrast between two adjacent tissues is likely to occur, making the delineation more difficult. In addition to these effects, it might also happen that some tissues or organs of interest share similar intensity levels with nearby regions, leading to a lack of strong edge or ridge information along the boundaries of the object. In these cases, structures of interest are very difficult to be separated from its surroundings. If the object to be segmented has a complex shape, this lack of contrast along the boundaries makes the segmentation even harder. Last, besides of the image information, higher level knowledge of anatomy and pathology is critical for medical image segmentation. Medical images have usually complex appearance due to the complexity of anatomic structures. Medical expertise is therefore required to understand and interpret the image so that the segmentation algorithms could meet the clinicians' needs.

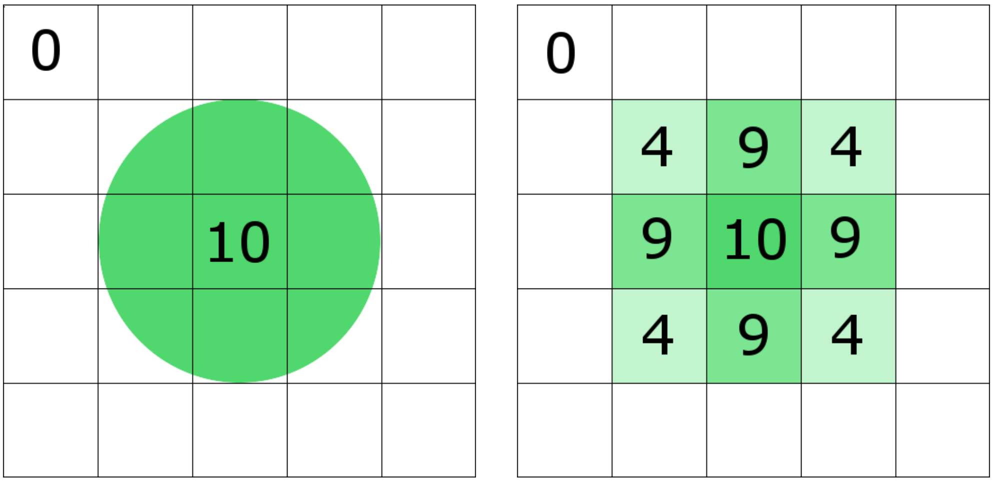


Figure 3.2: Partial volume effect caused by effects of finite voxel size when imaging a circle. The green area in the left image has value 10 and the white area has value 0. Imaging this circle with 9 voxels results in the right figure.

Despite these drawbacks, recent developments of medical imaging acquisition techniques, such as CT MRI have allowed to increase the resolution of images which have greatly assisted in clinical diagnosis. Nevertheless, these advances have not only significantly improved the resolution and information captured in the diverse image modalities, but also have led to an increase of the amount of data to be analyzed. Additionally, data complexity has been also affected. This increment in complexity has forced to medical technicians to process a large number of images with much more details.

3.3 Segmentation in neuroimaging

Initial approaches of brain segmentation on MRI focused on the classification of the brain into three main classes: white matter (WM), grey matter (GM) and cerebrospinal fluid (CSF) [34]. During the last two decades, the segmentation of the whole brain into the primary cerebrum tissues (i.e. CSF, GM, and WM) has been one of the core challenges of the neuroimaging community, leading to many publications. Nevertheless, it is still an active area of research [35, 36]. More recent methods include tumors and adjacent regions, such as necrotic areas [37]. Those methods are only based on signal intensity. However, segmentation of subcortical structures (i.e. OARs) can hardly be achieved based solely on signal intensity, due to the weak visible boundaries and similar intensity values between different subcortical structures. Consequently, additional information, such as prior shape, appearance and expected location, is therefore required to perform the segmentation.

Due to the crucial role of the hippocampus (HC) in learning and memory processes [38] and its role as biomarker for the diagnosis of neural diseases, such as Parkinson, dementia or Alzheimer [39], many methods have been published to (semi-) automatically segment the HC on MRI [40–55]. Among presented methods to segment the HC, atlas-based, statistical and deformable models have been typically employed.

Segmentation approaches of other brain structures, in addition to the HC, have been investigated. For instance, segmentation of corpus callosum has been approached by parametric [56] and geometric [57] deformable models. An active shape model method was employed in [58] to segment the mid brain on MR images. Other researchers have focused on a set of different subcortical and cerebellar brain structures instead, proposing several approaches: active shape and appearance models [59–64], atlas-based methods [65–69],deformable models [70–72] or machine learning approaches [73–76].

Notwithstanding, the number of publications focusing on segmentation of structures involved in the RTP is relatively lower. In addition, although good performance has been often reported for some of these structures, evaluation of proposed methods has been made on control and on several mental disorders patients, such as Schizophrenia or Alzheimer. Nevertheless, in brain cancer context, the presence of tumors may deform other structures and appear together with edema that changes intensity properties of the nearby region, making the segmentation more challenging.

There exist, however, a reduced number of approaches that have already attempted to segment some OARs and brain structures in patients undergoing radiotherapy [5, 6, 32, 77–80]. While for large structures results were often satisfactory, automatic segmentation of small structures were not sufficiently accurate for being usable in RTP in most cases. An atlas-based approach to segment the brainstem was validated in brain cancer context in [5]. In the work of [78], whilst segmentation of large structures was considerably suitable for RTP, optic chiasm and pituitary gland segmentations were totally unsuccessful. In other attempt to evaluate an automatic approach on a clinical environment, [6] also reported unsatisfactory results for small OARs such as the chiasm. Despite insufficient results reported on small OARs, previous works demonstrated that the introduction of automatic segmentation methods may be useful in a clinical context.

The objective of this chapter is to provide the reader with a summary of the current state of the art with regard to approaches to segment subcortical brain structures. As it has been reported in the previous section, a large number of techniques have been proposed over the years to segment specific subcortical structures in MRI. However, we are interested in those techniques which are typically applicable to subcortical brain structures in general. In the presented work, we mainly focus on minimally user-interactive methods -automatic or semi-automatic -, which are not tailored to one or few specific structures, but applicable in general. Thus, methods presented in this chapter can be divided into four main categories: *atlas-based methods, statistical models, deformable models and machine learning methods.*

3.4 Atlas-based segmentation methods

The transformation of brain MRI segmentation procedures from human expert to fully automatic methods can be witnessed by exploring the atlas-based methods. Segmentation by using atlas-based methods can be divided into the following main steps: atlas construction, registration between the atlases and the target image, and optionally atlas selection and label fusion (Figure 3.3).

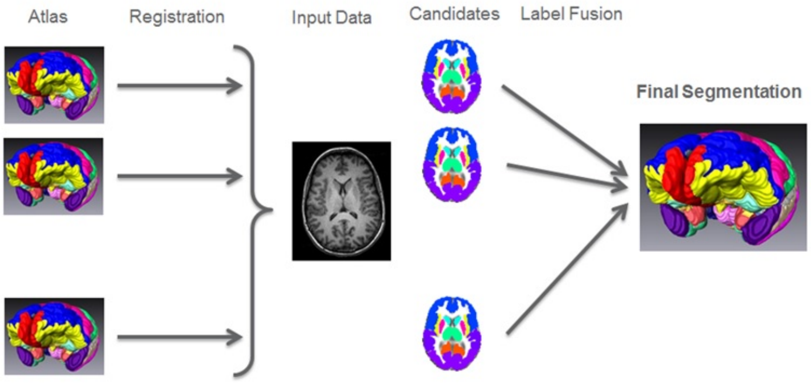


Figure 3.3: Typical atlas-based segmentation workflow where multiple atlases are employed.

3.4.1 Atlas build-up

First attempts at atlas construction of the human brain were based on a single subject. Here, a single atlas image is used to perform the segmentation [66].

This atlas, referred as topological, single-subject or deterministic atlas, is usually an image selected from a database to be representative of the dataset to be segmented, in terms of size, shape and intensity for instance. Particularly, for follow-up of patient's disease where segmentation of brain structures should be performed on longitudinal studies (i.e. at different time point along the treatment), the use of single-atlas based segmentation method to propagate segmented structures obtained at one time point to another time point is generally sufficient. However, in applications where no prior image of the patient can be used as atlas, the segmentation using single-atlas based methods of anatomical structures presenting wide variability between humans becomes challenging, and might lead to poor results.

To overcome the limitations encountered with single-atlas based method, multiple atlases can be used [5,44–46,49,50,54,65,67–69,81] . In this approach, multiple atlas images are selected from a database of images representative of the image to be segmented. Each atlas image is then registered to optimally fit the target image. Subsequently, using the deformation resulting from registration, the atlas labeled image is deformed. At this stage, multiple labeled images are fitted to the target image. At last, propagated labeled images are fused, providing the final segmentation. Beside the registration method used, performance of multi-atlas segmentation methods depends on: 1) the atlas building, 2) the atlas selection (Section 2.3), and 3) the label fusion method (Section 2.4) used. The major drawback of multi-atlas based segmentation methods remains the computation cost since it increases with the number of atlases selected.

A limitation of the multi-atlas based segmentation methods is that individual differences that occur in only a minority of the atlases could be averaged out. Hence, segmentation results might be biased, particularly for MRI scans presenting some pathologies. To address this issue, probabilistic atlases are used. This third category of atlases estimates a probabilistic model of the input images, either from a probabilistic atlas or a combination of topological atlases. For a more detailed explanation see the work of Cabezas et al. [82]

3.4.2 Image Registration

Image registration is a prerequisite to perform atlas-based segmentation. The registration process is used to spatially align an atlas A and the target image T. For our segmentation purpose, the registration process involved is necessarily based on non-rigid approaches to tackle inter-individual spatial variation. Various image registration methods exist and have been applied to many medical application domains. We refer the reader to the publications of Hill et al. [83] and Zitova and Flusser [84] for an overview of the image registration methods, regardless of particular application areas. A review of image registration approaches specifically used in brain imaging is available in the publication of Toga and Thompson [85]. The main contributions, advantages, and drawbacks of existing image registration methods are addressed.

3.4.3 Atlas selection

Normal individual variations in human brain structures present a significant challenge for atlas selection. Some studies demonstrated that, although the use of more than only one topological atlas improves the accuracy of the segmentation, it is not necessary to use all the cases in a dataset for a given query image [49, 54, 66–68, 86, 87]. Among the existing solutions to choose the best matching cases, the use of meta-information is the simplest case. In this solution, which can be also called population specific atlases, an average atlas is built for several population groups according to similar features, like gender or age. Although they represent the simplest solution, the use of meta-information has proved to be a powerful similarity criterion when used in multi-atlas segmentation [67]. However, this information may not be always available, requiring the use of similarity metrics to compare both atlas and target image.

Initially, the majority of published works used a single individual image randomly selected from the atlas dataset, where the selection criterion was not even mentioned. The optimal selection of a single template from the entire dataset during atlas-based segmentation and its influence in the segmentation accuracy was investigated in [86]. Han et al. [87] compared the selection of a single atlas against the propagation and fusion of their entire atlas database. In their work, the selection of the single atlas was based on the highest Mutual Information (MI) similarity between atlases and the target image after a global affine registration. Multi-atlas segmentation strategy significantly improved the accuracy of single-atlas based strategy, especially in those regions which represented higher dissimilarities between images. Additionally to MI, Sum of squared differences (SSD) or cross-correlation (CC) are often used as a similarity metric to select the closest atlas with respect to the target image.

Aljabar et al. [67] proved that using multi-atlas selection when segmenting subcortical brain structures improves the overlapping than when using random sets of atlases. In their work, a dataset of 275 atlases was used. As in [87], MI similarity was used to top-rank the atlases from the dataset. Then, the n top ranked atlases from the list were selected to be propagated to the target image by using a non-rigid registration. Mean DSC obtained by selecting the top-ranked atlases (0.854) was higher than the DSC obtained randomly selecting the atlases (0.811). This difference represents nearly 4% of improvement, demonstrating that the selection of a limited number of atlases which are more appropriate for the target image and prior to multi-atlas segmentation, would appear preferable to the fusion of an arbitrarily large number of atlases.

The inclusion in the label propagation step of atlases containing high dissimilarities with respect to the target image, may not make the segmentation more accurate, but contribute to a poorer result. Consequently, the proper selection of the atlases to include in the label propagation is a key step of the segmentation process.

3.4.4 Label fusion

Once the suitable atlases have been selected from the atlas dataset and labels propagated to the target image, information from transferred labels has to be combined to provide the final segmentation [44–46, 49, 50, 52, 54, 65, 67, 69, 81, 86, 88, 89]. This step is commonly referred as label fusion or classifier fusion.

Label fusion techniques known as best atlas and majority voting approach represent the simplest strategies to combine the propagated labels. In best atlas technique, after the registration step, the labels from the most similar atlas to the target image are propagated to yield the final segmentation. In majority voting method, votes for each propagated label are counted and the label receiving the most votes is chosen to produce the final segmentation [45, 65, 67]. Since majority voting assigns equal weights to different atlases, it makes a strong assumption that different atlases produce equally accurate segmentations for the target image.

To improve label fusion performance, recent work focuses on developing segmentation quality estimations based on local appearance similarity and assigning weights to the propagated labels. Thus, final segmentation is obtained by increasing the contribution of the atlases that are more similar to the target scan [44–46, 49, 50, 54, 66, 86]. Among previous weighted voting strategies, those that derive weights from local similarity between the atlas and target [44, 46, 49, 50], and thus allow the weights to vary spatially, have demonstrated to be a better solution in practice. Hence, each atlas contributes to the final solution according to how similar to the target they are. However, the computation of the weights is done independently for each atlas, and the fact that different atlases may produce similar label errors is not taken into account. This assumption can lead to labeling inaccuracies caused by replication or redundancy in the atlas dataset. To address this limitation, a solution for the label fusion problem was proposed [54]. In this work the weighted voting was formulated in terms of minimizing the total expectation of labeling error and the pairwise dependency between atlases was explicitly modeled as the joint probability of two atlases making a segmentation error at a voxel.

Hence, the dependencies among the atlases were taken into consideration, and the expected label error was reduced in the combined solution.

Another remarkable example of producing consensus segmentations, especially in the context of medical image processing, is the algorithm named Simultaneous Truth and Performance Level Estimation (STAPLE) [89]. STAPLE approach, instead of using an image similarity metric to derive the classifier performance, estimates the classifier performance parameters by comparing each classifier to a consensus, in an iterative manner according to the Expectation Maximization (EM) algorithm. In order to model miss registrations as part of the rater performance, a reformulation of STAPLE with a spatially varying rater performance model was introduced [88]. More recently, Cardoso et al. [52] extended the classical STAPLE approach by incorporating a spatially image similarity term into a STAPLE framework, enabling the characterization of both image similarity and human rater performance in a unified manner, which was called Similarity and Truth Estimation for Propagated Segmentations (STEPS). At last, a novel reformulation of the STAPLE framework from a non-local perspective, called Non-local Spatial STAPLE [69], was used as a label fusion algorithm [81].

3.4.5 Joint segmentation-registration

It is important to note that most atlas-based methods presented perform registration and segmentation sequentially. Nevertheless, there exist approaches that exploit complementary aspects of both problems to segment either several tissues [90–95] or tumors [96,97]. The idea of joining registration and segmentation has been utilized by boundary localization techniques using level set representation [57]. These methods relate both problems to each other by extending the definition of the shape to include its pose.

In the work of Yezzi et al. [90], a variational principle for achieving simultaneous registration and segmentation was presented. However, the registration step was limited to rigid motions. Another variational principle in a level-set based formulation was presented in the work of Paragios et al. [91] to jointly segment and register cardiac MRI data. A shape model based on a level set representation was constructed and used in an energy to force the evolving interface to rigidly align with the prior shape. The segmentation energy was separately involved as a boundary and region based energy model. In their work, again, the proposed formulation was limited to rigid motion. Departing from earlier methods, Wang et al. [93] proposed a unified variational principle where segmentation and non-rigid registration instead, were simultaneously achieved. Unlike previous approaches, their algorithm could accommodate for image pairs presenting a high variation on intensity distributions. Among other applications of this work, 3D hippocampal segmentation was presented. Wu et al. [95] also benefit from joint segmentation and registration to address the problem of segmentation of infant brains from subjects at different ages. In their work, tissue probability maps were separately estimated by using only training at the respective age. Probability maps were then employed as a good initialization to guide the level set segmentation. Some of these work have shown the improvements of coupling segmentation and registration with respect to their isolated use. Nevertheless, the use of this technique to segment some of the structures of interest for our particular problem is minimal, with very few published works [93].

3.4.6 Strengths and Weaknesses

Nearly all atlas-based techniques require some sort of image registration at the initial stages. That means that the success of the atlas propagation highly depends on the registration step. Regarding the creation of the atlases, they are relatively simply to build: any segmentation can be suitable for being an atlas.

The use of a single atlas to propagate segmented structures within a single patient (i.e. at different time point along the treatment for a given patient) is generally sufficient. However, in intra-patients situations presenting wide variability between humans the use of only one atlas might lead to unsatisfactory results. The use of more than one atlas improves segmentation quality in these situations. By increasing the number of atlases in the database, the method becomes more representative of the population and more robust when processing target images that can represent possible deviations. However, when working with multiple atlases, the key point is to determine which atlas must be used, that is not too different from the target image. To achieve this, some similarity metrics are used after the registration step and hence the choice of the closest atlas among all the others in the database. Alternatively to select the closest atlas to the target image, several atlases can be propagated, leading to multiple candidate segmentations that have to be merged at the end of the process. Merging of candidates is performed by label-fusion methods with the risk that these methods can generate organs with disconnected pieces, which is often hardly plausible from an anatomical point of view.

From a clinical perspective, recent clinical evaluations of the final segmentations still reveal the need of manual editing or correction of the automatic contours [98]. Additionally, the definition of an appropriate atlas or a set of appropriate atlases remains still an open question. Furthermore, no consensus exists on inclusion/exclusion rules of a given patient in a database, or in the numbers of patients to be included [67, 86]. Because of all these constraints, atlas-based segmentation techniques still suffer from a slow adoption by the physicians in clinical routine.

One of the main limitations of atlas-based methods is that the contours included in the atlases contain prior knowledge about organs pictured in the image which is not exploited. To perform the segmentation, these contours are merely deformed. As a consequence, most of the information conveyed by the contours, such as shape or appearance, remains implicit and likely underexploited. Statistical models are an alternative that address this issue by making a more explicit use of such prior information to assist the image segmentation. Unlike atlases, the images are not registered but the shapes and, sometimes, the appearance of the organ, are learned in order to be found in a target image.

3.5 Statistical models

Statistical models (SM) have become widely used in the field of computer vision and medical image segmentation over the past decade [48,58–64,99–113]. Basically, SMs use a priori shape information to learn the variation from a suitably annotated training set, and constrain the search space to only plausible instances defined by the trained model. The basic procedure of SM of shape and/or texture is as follows: 1) the vertices (control points) of a structure are modeled as a multivariate Gaussian distribution; 2) shape and texture are then parameterized in terms of the mean and eigenvectors of both the vertex coordinates and texture appearance; 3) new instances are constrained to a subspace of allowable shapes and textures, which are defined by the eigenvectors and their modes of variation. Consequently, if the dimensionality of the shape representation exceeds the size of the training data, the only permissible shapes and textures are linear combinations of the original training data.

3.5.1 Training Phase. Construction of the statistical model

3.5.1.1 Modelling the shape

Statistical shape model (SSM) construction basically consists in extracting the mean shape and a number of modes of variation from a collection of training samples to represent the possible shapes that the model is able to generate. Landmarks based method is a generic technique coined as Point Distribution Models (PDMs) by Cootes et al. [99], which has been extensively used in SSMs for surface representation. This method regularly distributes a set of points across the surface, which usually relies on high curvatures of boundaries (Figure 3.4. Images courtesy of [114]). However, they do not need to be placed at salient feature points as per the common definition of anatomical landmark, which is the reason of why they have also been referred as semi-landmarks. Among other shape representation models that have been recently used in medical image segmentation [108] we can identify medial models or skeletons, meshes, vibration modes of spherical meshes or the use of wavelets, for example.

Alignment of the training shape samples in a common coordinate frame is the first step to create the shape model. Once the samples are co-registered, a reduced number of modes of variation that best describes the variation observed are extracted, which is usually done by applying Principal Components Analysis (PCA) to the set of vectors describing the shapes [100]. PCA picks out the main axes of the cloud, and models only the first few, which account for the majority of the variation. Thus, any new instance of the shape can be modeled by the mean shape of the object and a combination of its modes of variations [99].

3.5.1.2 Modelling the appearance

As an extension of the statistical models of shape, the texture variability observed in the training set was included in the model, leading to appearance models (AMs) [102]. In this approach, in addition to the shape, the intensity variation seen in the training set is also modeled. As in the SSM, the variability observed in the training set is parameterized in terms of its mean and eigenvectors. Once the shape has been modeled (See section 3.1.1), the statistical model of the gray level appearance has to be built. For this purpose, sample images are warped based on the mean shape. Then, the intensity information from the shape-normalized image is sampled over the region covered by the mean shape. Different techniques to sample the intensity in the warped image can be found in the literature [108].

3.5.2 Segmentation Phase. Search algorithm

Once the SM has been created, it is important to define the strategy to search new instances of the model in the input images. This step consists essentially in finding the most accurate parameters of the statistical model that best define a new object. Active shape models(ASM) and active appearance models (AAM) are the most frequently employed constrained search approaches and are described below.

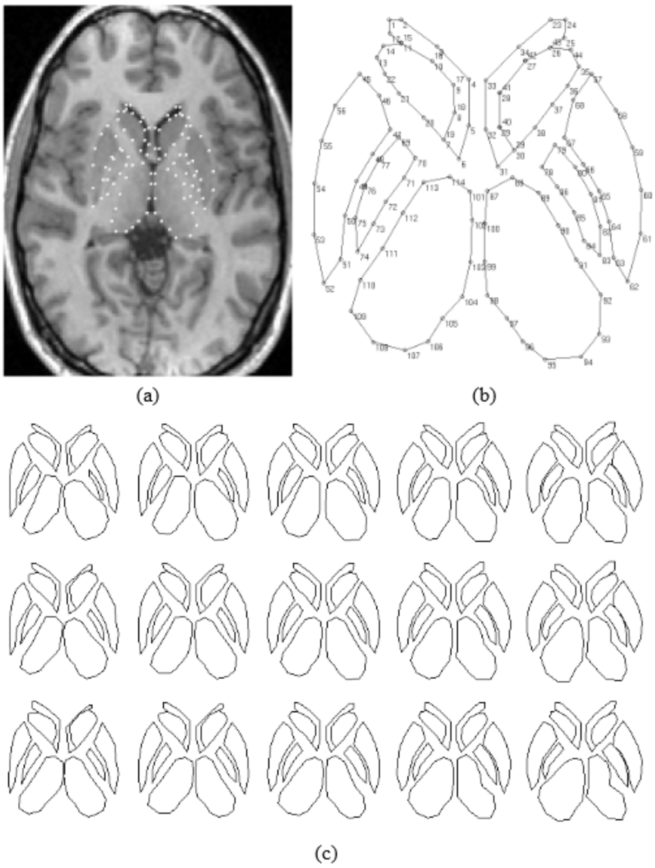


Figure 3.4: An example of constructing Point Distribution Models. (a) An MR brain image, transaxial slice, with 114 landmark points of deep neuroanatomical structures superimposed. (b) A 114-point shape model of 10 brain structures. (c) Effect of simultaneously varying the model's parameters corresponding to the first two largest eigenvalues (on a bi-dimensional grid)

3.5.2.1 Active Shape Model

Originally introduced by Cootes et al. [99,100], ASM is a successful technique to find shapes with known prior variability in input images. ASM has been widely used for segmentation in medical imaging [108], including segmentation of subcortical structures on brain [58,61,63,101,103–105,107,112,113]. It is based on a statistical shape model (SSM) to constrain the detected organ boundary to plausible shapes (i.e. shapes similar to those in the training data set). Given a coarse object initialization, an instance of the model can be fit to the input image by selecting a set of shape parameters defined in the training phase (see Section 3.1.1).

Original ASM method [100] was improved in [103] by using an adaptive gray-level AM based on local image features around the border of the object. Thus, landmarks points could be moved to better locations during the optimization process. To allow some relaxation in the shape instances fitted by the model, ASM can be combined with other methods, as in [104]. They employed a framework involving deformable templates constrained by statistical models and other expert prior knowledge. This approach was used to segment four brain structures: corpus callosum, ventricles, hippocampus and caudate nuclei. Most of the ASMs used in the literature are based on the assumption that the organs to segment are usually located on strong edges, which may lead to a final shape far from the actual shape model. Instead, [58] presented a novel method which was based on the combined use of ASM and Local Binary Patterns(LBP) as features for local appearance representations to segment the midbrain. In this way, segmentation performance was improved with respect to the ASM algorithm.

A major limitation of ASM is the size of the training set (especially in 3D), due to lack of representative data and time needed for model construction process. Hence, 3D ASMs tend to be restrictive in regard to the range of allowable shapes, over-constraining the deformation. Zhao et al. [105] overcame this limitation by using a partitioned representation of the ASM where, given a PDM, the mean mesh was partitioned into a group of small tiles, which were used to create the statistical model by applying the PCA over them. Other techniques focus on artificially enlarging the size of the training set. Koikkalainen et al. [106] concluded that the two best enlargement techniques were the non-rigid movement technique and the technique that combines PCA and a finite element model.

3.5.2.2 Active Appearance Model

The active appearance model (AAM) is an extension of the ASM that, apart from the shape, models both the appearance and the relationship between shape and appearance of the object [102]. Since the purpose of this review is to give a view about the use of these methods in medical image segmentation (especially of the subcortical structures on MRI), and not to enter into detail in the mathematical foundations of each methods, we encourage the readers to review a detailed description of the algorithm in [102].

Initially, Cootes et al. [59] demonstrated the application of 2D AAMs on finding structures in brain MR images. Nevertheless, they are not suitable for 3D images in their primary form because of the underlying shape representation (i.e. PDM) that becomes impractical in 3D. Some approaches extended them to higher dimension by using non-linear registration algorithms for the automatic creation of a 3D-AAM. Duchesne et al. [60] segmented medial temporal lobe structures by including nonlinear registration vector fields into a 3D warp distribution model.

However, a number of considerations have to be taken into account in adapting a generic AAM approach to a specific task. Babalola et al. [109] built AAMs of some subcortical structures using groupwise registration to establish correspondences, i.e. to initialize the composite model within the new image. To build the AAMs, the intensities along vectors normal to the surface of the structures were sampled, which is known as profile AAM. In [62], the proposed approach used a global AAM to find an approximate position of all the structures in the brain. Once the coarse localization was found, shape and location of each structure were refined by using a set of AAMs individually trained for each of the structures. Although the probability of object occupancy could be derived from the training set, they demonstrated that the use of simple regressors at each voxel based on the pattern of grey level intensities nearby provided better results.

3.5.2.3 Initialization

Most of the methods that aim to locate a SSM in a new input image use a local search optimization process. So, they need to be initialized near the structure of interest, so that the model boundaries fall in the close vicinity of object boundaries in the image. Straightforward solution for the initialization problem is human-interaction. In some cases, it is sufficient to roughly align the mean shape with the input data, whereas in other cases, it is preferred to use a small number of points to guide the segmentation process [103]. Alternatively, more robust techniques can be used to initialize the model in the image [109–111]. Nevertheless, the automatic methods can be slow, especially when they work with 3D images.

3.5.3 Strengths and Weaknesses

Unlike atlas-based segmentation methods, statistical models require a learning model. Mean shapes, textures and their modes of variations which define this model are learned from the training set. If the number of samples used to build the learning model is not sufficient, there is a significant risk to overfit the shape or the appearance. If the number of images used to build the model is low, there is a non-negligible risk to overfit the shape and/or the appearance. Overfitting arises when the learned model is too specific to the training set and is not able to acceptable fit unseen instances. Then, it performs well on the training samples but its performance is quite poor when dealing with new examples. Additionally, if some noise along the shapes is learned in the model, robustness when segmenting target images will be also affected.

When utilizing the ASM, during the optimization process, the intensity model and the shape model are applied alternatively. First, candidate target points in the neighborhood of each landmark point are search. And second, a new ASM shape is fit through these points. This procedure is repeated iteratively until convergence. The fact that the shape model may be deceived if the gray-level appearance model does not select a proper landmark makes ASM methods sensitive to local optima.

Because of target points are searched in a local constrained vicinity of the current estimation for each landmark location, a sufficiently accurate initialization needs to be provided in order to make the model converge to the proper shape. Therefore, for both ASM and AAM, the search of the shape and/or appearance requires an initialization. It can be provided either by direct human-interaction or by automatic techniques, which might result too slow. If the initial position is too distant from the searched object, in terms of translation, rotation or scale, this can lead to poor object identification.

3.6 Deformable models

The term ”deformable model” (DM) was pioneered by Terzopoulos et al. [115] to refer to curves or surfaces, defined in the image domain, and which are deformed under the influence of internal and external forces. Internal forces are related with the curve features and try to keep the model smooth during the deformation process. In the other hand, external forces are the responsible of attracting the model toward features of the structure of interest, and are related with the image features of the adjacent regions to the curve. Hence, DM tackles the segmentation problem by considering an object boundary as a single, connected structure, and exploiting a priori knowledge of object shape and inherent smoothness [115]. Although DM were originally developed to provide solutions for computer vision applications to natural scenes and computer graphics problems, their applicability in medical image segmentation has already been proven [116]. An example of using deformable models to segment the corpus callosum is shown in Figure 3.5 (Images courtesy of [117]).

According to the type of shape representation used to define the model, DM methods can be categorized in: parametric or explicit deformable models [56, 70, 118–120] and geometric or implicit deformable models [40, 51, 57, 71, 72, 121–125].

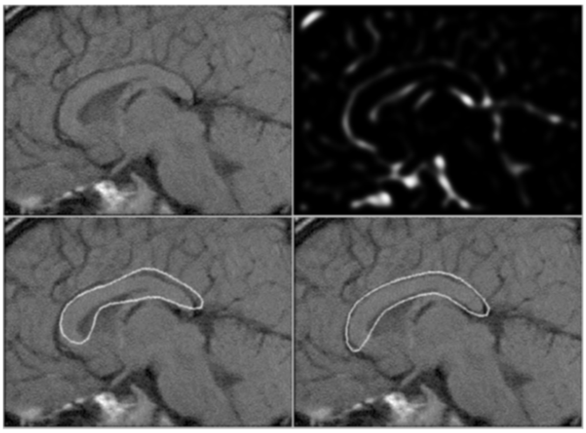


Figure 3.5: Segmenting the corpus callosum from an MR midbrain sagittal image using a deformable Fourier model. Top left: MR image (146 x 106). Top right: positive magnitude of the Laplacian of the Gaussian (γ= 2.2) Bottom left: initial contour (six harmonics). Bottom right: final contour on the corpus callosum of the brain.

3.6.1 Parametric deformable models

The first parametric model used in image segmentation found in the literature was originally introduced by Kass et al. [118], coined with the name of ?snakes?. It was proposed as an interactive method where, because of its limitations, initial contours must be placed within the vicinity of object boundaries. First, the energy of the contour depends on its spatial positioning and changes along the shape. Sensitivity to initial location obliges the contour to be placed close to the object boundary, leading to failure in case of improper initialization. Second, the presence of noise may cause the contour to be attracted by a local minimum and get stuck in a location that might not correspond with the ground truth. To overcome these limitations different approaches have been proposed [116, 119]. The method presented in [119] provides different mechanisms to enable the contour topology to change during the deformation process. In [116], an extensive study of DM and different types of external forces was presented.

Regarding the segmentation of subcortical structures, parametric DM have been recently employed to perform the segmentation, in combination with other approaches [56, 70, 120]. Ada-boosted algorithm was used in [120] to detect brainstem and cerebellum candidate areas, followed by an active contour model to provide the final boundaries. An extension of natural snakes was proposed in [70], where desired properties of physical models were combined with Fourier parameterizations of shapes representations and their shape variability to segment the corpus callosum. In [56], the application of genetic algorithms to DM was explored in the task of corpus callosum segmentation. In this approach, genetic algorithms were propose to reduce typical deformable model weaknesses pertaining to model initialization, pose estimation and local minima, through the simultaneous evolution of a large number of models.

3.6.2 Geometric deformable models

One of the main drawbacks of parametric DM is the difficulty of naturally handling topological changes for the splitting and merging of contours, restricting severely the degree of topological adaptability of the model. To introduce topological flexibility, geometric DM have been implicitly implemented by using the level set algorithm developed by Osher and Sethian [121]. These models are formulated as evolving contours or surfaces, usually called fronts, which define the level set of some higher-dimensional surface over the image domain.

Generally, image gray level based methods face difficult challenges such as poor image contrast, noise, and diffuse or even missing boundaries, especially for certain subcortical structures. In most of these situations, the use of prior model based algorithms can solve these issues. The method proposed in [122] used a systematic approach to determine a boundary of an object as well as the correspondence of boundary points to a model by constructing a statistical model of shape variation. Ghanei et al. [40] used a deformable contour technique to customize a balloon model to the subjects' hippocampus. In order to avoid local minima due to mismatches between model edge and multiple edges in the image, their technique incorporates statistical information about the possible range of allowable shapes for a given structure. Geodesic active contours were extended in [57] by incorporating shape information into the evolution process. PCA and level set functions of the object boundaries were employed to form a statistical shape model from the training set. The segmenting curves evolved according to image gradients and a maximum a posteriori (MAP) estimated the shape and pose.

The use of level set methods to formulate the segmentation problem has been reported to increase the capture range of DM and constrain the deformation through the incorporation of some prior shape information. Because of these advantages geometric DMs have been extensively used to carry out the segmentation task of brain subcortical structures [40, 57, 71, 72, 122–125].

In some situations, texture information is also required to constrain the deformation on the contours. As a consequence, statistical models of both shape and texture are used in addition to only shape prior based segmentation methods [59, 102]. The modeled structure can be located by finding the parameters, which minimize the difference between the synthesized model image and the target image in conjunction with the statistical model of the shape based on landmark points and texture.

3.6.3 Strengths and Weaknesses

Contrary to statistical models, no training or previous knowledge is required by deformable models. These models can evolve to fit into the desired shape, showing more flexibility than other methods. Nevertheless, the definition of stopping criteria might become hard to achieve, and it depends on the characteristics of the problem.

Parametric deformable models have been successfully employed in a broad range of applications and problems. An important property of this kind of representation is its capability to represent boundaries at a sub-grid resolution, which is essential in the segmentation of thin structures. However, they present two main limitations. First, if variation in size and shape between the initial model and the target object are substantial, the model must be reparameterized dynamically to faithfully recover the boundary of the object. The second limitation is related with the complications that they present to deal with topological changes, such as splitting or merging model parts. This property is useful to recover either multiple objects or an object with unknown topology. Geometric models, however, provide an elegant solution to address these main limitations of parametric models. Due to these models are based on curve evolution theory and the level set method, curves and surfaces evolve independently of the parameterization. Evolving curves and surfaces can therefore be represented implicitly as a level set of a higher-dimensional function, resulting in automatic handling of topological transitions.

Although topological adaptation can be useful in many applications, it can sometimes lead to undesirable results. Geometric deformable models may generate shapes that have inconsistent topology with respect to the actual object, when applied to noisy images with significant boundary gaps. In these situations, the significance of ensuring a correct topology is often a necessary condition for many subsequent applications. Parametric deformable models are better suited to these applications because of their strict control on topology. Additionally, in practice, design of parametric deformable models is more straightforward because of its discrete representation rather than a continuous curve or surface, like in the geometric deformable models. A common disadvantage that share both geometric and parametric models is that their robustness is limited to specific type of images. Suitable images to apply any of the deformable models here presented must provide sufficient edge or region-based information for an explicit modeling in a deterministic or probabilistic manner with parametric assumptions. As a consequence, traditional deformable models generally fail to segment images with significant intensity inhomogeneity and/or poor contrast.

3.7 Machine learning methods

Machine Learning (ML) techniques have been extensively used in the MRI analysis domain almost since its creation. Artificial Neural Networks (ANN), or Support Vector Machines (SVM), are among the most popular learning methods used not only for segmentation of brain anatomical structures [42, 43, 47, 73–76, 126–128] ,but also for tumors classification [129–131] or automatic diagnosis [132]. Although to a lesser extent, some brain structures others than WM, GM and CSF have also benefit from the use of some other machine learning approaches, such as k-Nearest Neighbors (KNN) [133–135]. Such supervised learning based segmentation methods first extract image features with information often richer than intensity information alone, and then construct a classification model based on the image features using supervised learning algorithms. We will first review typical features utilized in supervised learning based classification schemes (Section 3.7.1). Next, in section 3.7.2, some of the most common machine learning techniques employed to segment brain structures are presented.

3.7.1 Features used in segmentation

Among all possible information that can be extracted to segment brain structures in medical images, intensity-based, probability-based and spatial information are the most commonly employed features. They represent the simplest cases of features, in terms of complexity.

3.7.1.1 Intensity Features

Intensity features exploit intensity information of a voxel and appearance of its vicinity. Researchers have extracted neighborhood information in several ways. In its simplest representation, square patches around a given pixel are used in 2D, with typical patch size values ranging from 3 to 9 pixels(Figure 3.6). To catch texture appearance of a voxel and its neighbors cubic patches of different sizes -usually of size 3,5 or 7- are extracted in 3D (Figure 3.7). Extracting such cubic patches represents to have a subset of 27, 125 and 343 intensity values, respectively. These amounts of voxels, however, become sometimes very expensive and impractical, especially for large structures, where a larger number of instances is required in training. To offer a "cheaper" solution that still catches information as far away as these cubic patches, some works have proposed to use crosses orthogonal to the voxel under examination instead. In this way, capturing texture appearance in a radius of size 2 from the voxel v, for example, will lead to a total of 12 voxels, instead of 125 in the case of the cubic patch of size 5, while having the same scope. As alternative to square and cubic intensity patches and crosses, gradient direction has been used to capture relevant information of texture appearance. Here, intensity values along the gradient descents are used to characterize the voxel v and its surroundings. Taking intensity values along the maximum gradient direction from a few voxels from inside to outside has a distinct advantage over using neighbor intensity values based on a rectilinear coordinate system.

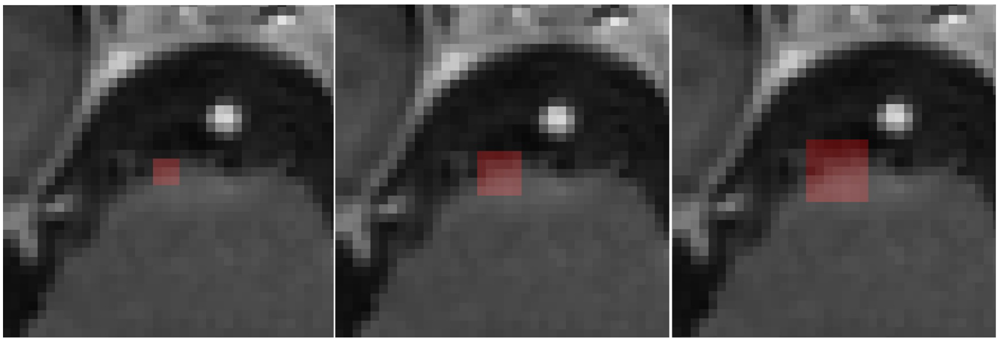


Figure 3.6: Intensity patches commonly used in 2D. Patch sizes are 3x3, 5x5 and 7x7 from left to right.

Image intensity has been largely used to segment objects in medical images. Indeed, it represents the fundamental feature utilized by the algorithms pioneering the use of ANN in the area of tissue classification [136,137] . Nevertheless, image intensity information individually is not good enough for distinguishing different brain structures since most of them share similar intensity patterns in MRI. To address such a problem, in learning based segmentation methods, more discriminative features are often extracted from MRI. In addition to image intensity values, which we will denote as IIV onwards, of voxels and their neighborhood, probabilistic and spatial information is often used.

3.7.1.2 Probability based Features

Probability based features are spatial probabilistic distribution maps for the different structures. They analyze the likelihood of a voxel to belong to a determined structure. The higher the value of a structure at a given location, the more likely the voxel at that location to be the structure. Probability maps generated for machine learning based systems can be seen like a sort of probabilistic atlases, but with more relaxed registration constraints. Labeled patients in the training set are employed to build a map of probabilities. To

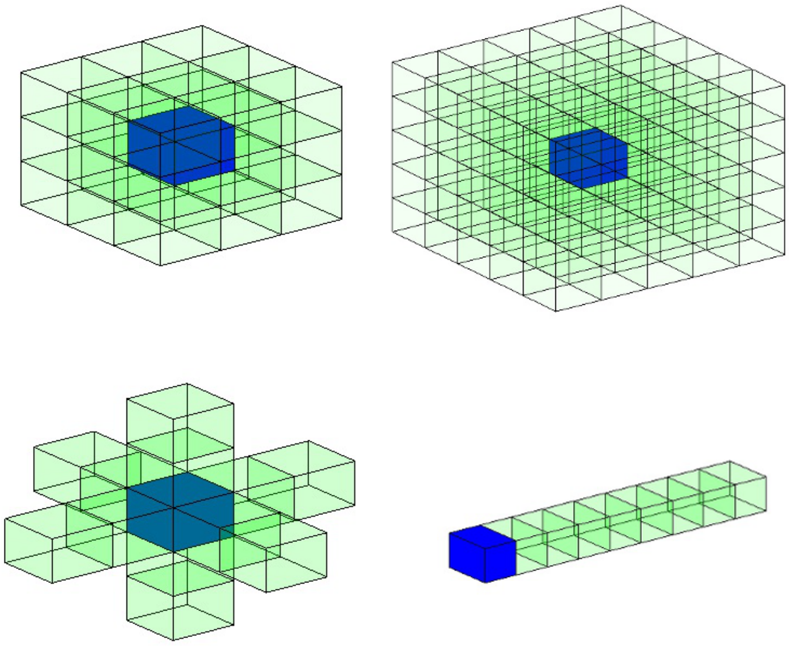


Figure 3.7: Intensity configurations commonly used in 3D. In blue is painted the center voxel under examination and in green its neighboring voxels.

ensure the probabilities on the map make sense, labels must be referred to the same reference system. To do so, an alignment of both MRI images and labels is required. Once all the patients have been aligned, labels are added to a common volume, creating the probability map (Figure 3.8).

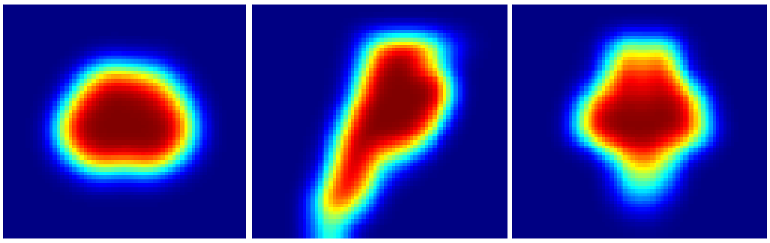


Figure 3.8: Brainstem probability map created from the training set.

3.7.1.3 Spatial based Features

Apart from image intensity and probability information, spatial knowledge of the voxel under examination can be employed. Although Cartesian coordinates (x,y,z) are frequently exploited, spherical coordinates (r, θ, φ) have also been used to capture the spatial information [138].

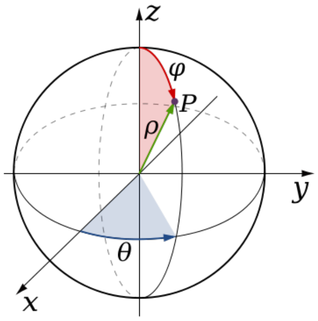


Figure 3.9: Cartesian and spherical coordinates.

Spatial information can aid in classification in several ways. First, the number of possible anatomical classes, such as the brainstem or the optic chiasm, at a given global position in the brain as specified by an atlas coordinate is often relatively small. Second, neuroanatomical structures occur in a characteristic spatial pattern relative to one another. For instance, taking the amygdala as example, it is anterior and superior to the hippocampus. And third, many tissue classes, such as gray or white matter, have spatially heterogeneous MRI intensity properties that vary in a spatially predictable fashion.

3.7.2 Learning Methods

The goal of many learning algorithms is to search a family of functions so as to identify one member of the mentioned family which minimizes a training criterion. The selection of this family of functions, as well as how members of that family are parameterized is of vital importance. Even though there is no universally optimal choice of parametrization of a family of functions (also called architecture) it might happen that some architectures are appropriate, or not, for a broad class of learning tasks and data distributions. Different architectures have different peculiarities that can be appropriate or not, depending on the learning task we are interested in. One of these characteristics, which has prompted a lot of interest in the research community in latest years, is the depth of the architecture. Depth corresponds to the number of hidden and output layers in the case of multilayer neural networks, which will be later introduced. Typical shallow neural networks are built of one to three hidden layers. In the case of support vector machines, for instance, depth is considered to be equal to two [139]. These architectures composed by very few layers are known as shallow architectures. Multilayer neural networks and support vector machines are among the most employed shallow architectures to perform classification. On the other hand, there are some other methods that, although they represent the simplest form of machine learning, have been employed to segment brain structures: KNN. Even though there exist some other methods inside this category that have been employed to segment either tumors or the brain in its primary classes, their contribution to segment critical brain structures has been marginal. Therefore, they are not considered in this review.

3.7.2.1 K-Nearest neighbors

K-Nearest neighbors (KNN) classification is based on the assignment of samples, i.e. image voxels, to a class, i.e. tissue type, by a search for samples in a learning set with approximately the same features. The learning set, generated from the labeled voxels, is entered into the feature space according to the feature values of its samples. A new image voxel is classified by inserting it in the feature space and further inspection of the K learning samples which are closest in a distance measure d to it. Then the tissue label is assigned to the target voxel based on a voting strategy among the tissues assigned to the K training voxels [140]. A common way to do this is to assign the most frequent class among the K neighbors to this voxel.

Although KNN it is very simple and easy to understand it has been successfully employed for segmentation on brain structures on MRI [133–135]. Anbeek et .al [133] proposed an automatic approach based on KNN and multi-parametric MRI for probabilistic segmentation of eight tissue classes in neonatal brains. Among evaluated structures, brainstem and cerebellum were included. Intensity values from the different MRI modalities were employed as features: T1- and T2-weighted ($T1_w$ and $T2_w$, respectively). In addition to intensity values, spatial information for each voxel was also used. Thus, each voxel was described with intensity and spatial features. Based on these features, each voxel was assigned to one of the eight tissue classes using a KNN-based classifier. Another attempt to segment brain structures by employing multi-parametric MRI in a KNN-based classifier was presented in [134]. In addition to $T1_w$ and $T2_w$ sequences, Proton Density weighted (PD_w) images were used to generate the voxel intensity information. As in [133], authors including spatial information into the features array by employing the x, y and z coordinates of the voxel under examination. More recently, Larobina et al. [135] investigated the feasibility of KNN to segment the four subcortical brain structures: caudate, thalamus, pallidum, and putamen. As in previous works, a combination of intensity and spatial-based information is employed to classify voxels. In their work, multispectral MRI from two studies were used.

While the first group was composed by $T1_w$, $T2_w$ and PD_w, the second group contained $T1_w$, $T2_w$ and FLAIR images. Additionally, they proposed the use of atlas-guided training as effective way to automatically define a representative and reliable training dataset, giving supervised methods the chance to successfully segment brain MRI images without the need for user interaction.

One of the main advantages of KNN-based classifier is that it is a very simple classifier that works well on basic recognition problems. Due to the nature of its mathematical background, training is performed relatively fast. Nevertheless, it does not learn anything from the training data and simply uses the training data itself for classification. To predict the label of a new instance the KNN algorithm will find the K closest neighbors to the new instance from the training data, the predicted class label will then be set as the most common label among the K closest neighboring points. The main disadvantage of this approach is that the algorithm must compute the distance and sort all the training data at each prediction, which can be slow if there are a large number of training examples. Another disadvantage of not learning anything from the training data, is that it can result in a model not generalizing well and also not being robust to noisy data. Further, changing K may affect the resulting predicted class label. In addition, if the available training set is small there exist a high risk of overfitting. Another drawback of KNN is that prediction accuracy can quickly degrade when number of attributes grows. Computation cost is very high because distance for each query instance to all training samples must be computed.

3.7.2.2 Artificial neural networks

An artificial neural network (ANN) represents an information processing system containing a large number of interconnected individual processing components, i.e. neurons. Motivated by the way the human brain processes input information, neurons work together in a distributed manner inside each network to learn from the input knowledge, process such information and generate a meaningful response. Each neuron n inside the network processes the input through the use of its own weight w_n, a bias value b_n, and a transfer function which takes the sum of w_n and b_n. Depending on the transfer function selected and the way the neurons are connected, distinct neural networks can be constructed.

Because of their efficacy in solving optimization problems, ANN have been integrated in segmentation algorithms to define subcortical structures [42, 73, 74, 76, 126, 128]. In the method proposed in [42], grey-level dilated and eroded versions of the MR T1 and T2-weighted images were used to minimize leaking from the HC to surrounding tissue combined with possible foreground tissue. An ANN was applied to a manually selected bounding box, which result was used as an initial segmentation and then used as input of the grey-level morphology-based algorithm. Magnotta et al. [73] used a three-layer ANN to segment caudate, putamen and whole brain. The ANN was trained using a standard back-propagation algorithm and a piecewise linear registration was used to define an atlas space to generate a probability map which was used as input feature of the ANN. This approach was later employed by [126] and extended by [74] through the incorporation of a landmark registration to segment the cerebellar regions. Based on the success of applying ANN approaches to segment cerebellar regions by incorporating a higher dimensional transformation, Powel et al. [76] extended the initial algorithm of [73] to use a high dimensional intensity-based transform. Further, they compared the use of ANN with SVM, as well as with more classical approaches such as single-atlas segmentation and probability based segmentation. In [128], a two-stage method to segment brain structures was presented, where geometric moment invariants (GMI) were used to improve the differentiation between the brain regions. In the first stage, GMI were used along voxel intensity values as an input feature and a signed distance function of a desired structure as an output of the network. To represent the brain structures, the GMI were employed in 8 different scales, using one ANN for each of the scales. In the second stage, the network was employed as a classifier and not as a function approximator.

Some limitations must be taken into account when ANN are employed. Their performance strongly depends on the training set, achieving good results only in those structures for which a suitable training can be developed. This may limit their value with inherently difficult structures that human beings have difficulty delineating reliably, such as the thalamus [73]. As a consequence, ANN must be well designed, and different types of ANN may require specific training data set development, depending on the structure-identification task.

3.7.2.3 Support vector machine

Another widely employed ML system, which also represents a state-of-the-art classifier, is Support Vector Machines (SVM). It was originally proposed by Vapnik [141] and [142] for binary classification. In contrast with other machine learning approaches like artificial neural network which aims at reducing empirical risk, SVM implements the structural risk minimization (SRM) that minimizes the upper bound of generation error.

Support vector machines (SVM), often called kernel-based methods, have been extensively studied and applied to several pattern classification and function approximation problems. Basically, the main idea behind SVM is to find the largest margin hyperplane that separates two classes. The minimal distance from the separating hyperplane to the closest training example is called margin. Thus, the optimal hyperplane is the one providing the maximal margin, which represents the largest separation between the classes. This will be the line such that the distances from the closest point in each of the two groups will be farthest away. The training samples that lie on the margin are referred as support vectors, and conceptually are the most difficult data points to classify. Therefore, support vectors define the location of the separating hyperplane, being located at the boundary of their respective classes. By employing kernel transformations to map the objects from their original space into a higher dimensional *feature space* [143], SVM can separate objects which are not linearly separable (Figure 3.10). Their good generalization ability and their capability to successfully classify non-linearly separable data have led to a growing interest on them for classification problems.

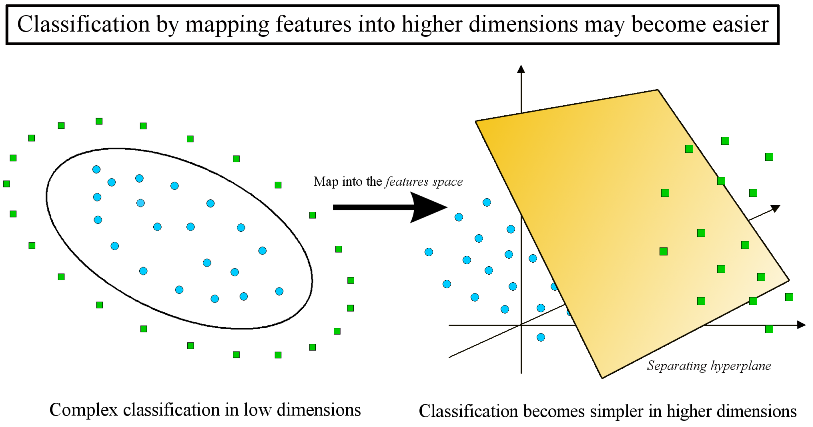

Classification by mapping features into higher dimensions may become easier
Map into the *features space*
Separating hyperplane
Complex classification in low dimensions
Classification becomes simpler in higher dimensions

Figure 3.10: Effect of the kernel transformation. Data is not linearly separable in (a). Mapping features into a higher dimensionality (b) may make the classification possible.

Support vector machines is a non-probabilistic supervised binary classifier that learns a model which represents the instances as points in space, mapped in such a way that instances of different classes are separated by a hyperplane in a high dimensional space. However, if the dataset is not linearly separable in that space the hyperplane will fail in classifying properly. This can be solved by mapping the dataset instances into a higher dimensional space using a kernel function, thus making easier the dataset division

Support vector machine represent one of the latest and most successful statistical pattern classifiers. It has received a lot of attention from the machine learning and pattern recognition community. Although SVM approaches have been mainly employed for brain tumor recognition [129–131] in the field of medical image classification, recent works have also used them for tissue classification [127] and segmentation of anatomical human brain structures [43, 47, 75, 76].

The growing interest on SVM for classification problems lies in its good generalization ability and its capability to successfully classify non-linearly separable data. First, SVM attempts to maximize the separation margin - i.e., hyperplane- between classes, so the generalization performance does not drop significantly even when the training data are limited. Second, by employing kernel transformations to map the objects from their original space into a higher dimensional feature space [143], SVM can separate objects which are not linearly separable. Moreover, they can accurately combine many features to find the optimal hyperplane. Hence, as can be seen, SVM globally and explicitly maximize the margin while minimizing the number of wrongly classified examples, using any desired linear or non-linear hypersurface.

Powell et al. [76] compared the performance of ANN and SVM when segmenting subcortical (caudate, putamen, thalamus and hippocampus) and cerebellar brain structures. In their study the same input vector was used in both machine learning approaches, which was composed by the following features: probability information, spherical coordinates, area iris values, and signal intensity along the image gradient. Although results obtained where very similar, ANN based segmentation approach slightly outperformed SVM. However, their employed a reduced number of brains to test (only 5 brains), and 25 manually selected features, which means that generalization to other datasets was not guarantee. PCA was used in [75] to reduce the size of the input training pool, followed by a SVM classification to identify statistical differences in the hippocampus. In this work, in addition to the input features used in [76], geodesic image transform map was added as input vector of the SVM. However, selection of proper discriminative features is not a trivial task, which has already been explored in the SVM domain. To overcome this problem, AdaBoost algorithm was combined with a SVM formulation [47]. AdaBoost was used in a first stage to select the features that most accurately span the classification problem. Then, SVM fused the selected features together to create the final classification. Furthermore, they compared four automated methods for hippocampal segmentation using different machine learning algorithms: hierarchical AdaBoost, SVM with manual feature selection, hierarchical SVM with automated feature selection (Ada-SVM), and a publicly available brain segmentation package (FreeSurfer). In their proposed study, they evaluated the benefits of combining AdaBoost and SVM approaches sequentially.

3.8 Discussion

Generally, none of the presented methods can singly handle brain subcortical structures segmentation with the presence of brain lesions. Typically, methods discussed in this survey rely on the existent information in a training set. However, subjects presenting brain lesions are not usually representative for a large set of patients, because of lesions may strongly differ and produce random deformations on the subcortical structures. As a consequence, they are not included in the training stage and the deformations on the structures caused by the lesion cannot be therefore modeled. A summary of referenced methods to segment subcortical structures is presented in Table 3.1. Additionally, details of the validation process for these methods are presented in tables 3.2 and 3.3. Definition and description of a validation process is of vital importance to evaluate segmentation methods in medical images. Nevertheless, since this process is not standardized there exist a lot of works that do not fully present all these details. In these two tables, we did our best to try to summarize all this important information.

Model based approaches, such as atlas or statistical models trend to perform reasonably well when there is no high anatomical deviation between the training set and the input case to analyze. Nevertheless, these approaches might completely fail if shape variability is not properly modeled, which often occurs in the presence of brain lesions. Additionally to the shape variability, registration plays an important role in atlas-based approaches. Registrations with large initial dissimilarity in shape between the atlases and the target might not be handled properly. This can lead to inappropriately weights when there are initially large shapes differences resulting in incorrect image correspondences established by the atlas registration. In the other hand, in statistical model approaches, which are only capable of generating a plausible range of shapes, the presence of a tumor might deform a determined structure to an unpredictable shape. This will cause the failure of SM approaches, because of their incapability to generate new unknown shapes which considerably differs from the shapes in the training set.

In the context of SMs, PCA was originally used in a framework called Active Shape Model(ASM) [100] and has become a standard technique used for shape analysis in segmentation tasks, and the preferred methodology when trying to fit a model into new image data. Compared to ASM, AAM makes an excessive usage of the memory when it creates the 3D texture model, and the implementation of ASM is relatively easier than the AAM implementation. While ASMs search around the current location and along profiles, AAMs only examine the image under its current area of interest, allowing the ASMs to generally have a larger capture range. However, the use of information solely around the model points makes that ASMs may be less reliable, since they do not profit from all texture information available across a structure, unlike AAM. Another interest advantage of the AAMs reported by [59] is related with the number of landmarks required to build a statistical model. Compared to the ASMs, AAMs can build a convincing model with a relatively small number of landmarks, since any extra shape variation may be encoded by additional modes of the texture model. Consequently, although the ASM is faster and achieves more accurate feature point location than the AAM, the AAM gives a better match to the image texture, due to it explicitly minimizes texture errors. Furthermore, ASM is less powerful in detecting the global minima and may converge to a local minimum due to multiple nearby edges in the image. These situations make AAM usually more robust than ASM. Although the main advantage of using PCA in SMs is to constraint the segmentation task to the space spanned by the eigenvectors and their modes of variation, it has two major limitations. First, the deformable shapes that can be modeled are often very restricted. Secondly, finer local variations of the shape model are not usually encoded in these eigenvectors. Consequently, new instances containing these small variations will not be properly fitted in the model instance.

Contrary to statistical models, DM provide flexibility and do not require explicit training, though they are sensitive to initialization and noise. SMs may lead to greater robustness, however they are more rigid than DM and may be over-constrained, not generalizing well to the unsampled population, particularly for small amounts of training data relative to the dimensionality. This situation can appear on new input examples with pathologies, lesions or presenting high variance, different from the training set. Models having local priors similar to DM formulation do not have this problem. They will easily deform to highly complex shapes found in the unseen image. Hence, many methods attempt to find a balance between the flexibility of the DM and the strict shape constraints of the SM by fusing learned shape constraints with the deformable model.

Notwithstanding, some main limitations have to be taken into account when working with generic parametric DM. First, if the stopping criterion is not defined properly, or boundaries of the structures are noisy, DM may get stuck in a local minimum which does not correspond to the desired boundary. Second, in situations where the initial model and the desired object boundary differ greatly in size and shape, the model must be reparameterized dynamically to faithfully recover the object boundary. Methods for reparameterization in 2D are usually straightforward and require moderate computational overhead. However, reparameterization in 3D requires complicated and computationally expensive methods. Further, it has difficulties when dealing with topological adaptation, caused by the fact that a new parameterization must be constructed whenever the topology change occurs, which may require sophisticated schemes. This issue can be overcome by using LSs. Moreover, as DM represent a local search, they must be initialized near the structure of interest.

By introducing machine learning methods, algorithms developed for medical image processing often become more intelligent than conventional techniques. Improvements in the resulting relative overlaps came from the application of the machine learning methods including ANN and SVM [76]. A comparison done in this work between four methods (template based, probabilistic atlas, ANN and SVM) showed that machine learning algorithms outperformed the template and probabilistic-based methods when comparing the relative overlap. There was also little disparity between the ANN and SVM based segmentation algorithms. ANN training took significantly longer than SVM training but can be applied more quickly to segment the regions of interest. It was reported that it took a day to train an ANN for the classification of only one structure from the others even though a random sampled data was used instead of the whole dataset.

Machine learning techniques have therefore demonstrated to outperform other, more traditional, approaches in segmenting brain structures. Recent developments of medical imaging acquisition techniques have led to an increase of complexity on the analysis of images. This brings new challenges where the analysis of large amount of data is compelled. On this context, we believe that machine learning techniques suit perfectly to deal with these new challenges.

However, a new area of Machine Learning has recently emerged with the intention of moving machine learning closer to one of its original purposes: Artificial Intelligence. This area is known as deep learning. Recent progress on using deep networks for image recognition, speech recognition, and some other applications has shown that they currently provide the best solutions to many of these problems. Therefore, we are going to consider the use of deep learning to address the problem of segmentation of brain structures in radiation therapy. Next chapter will introduce the reader in the context of deep learning and its use in this dissertation.

Method	Ref	Structures	Image Modality
Single Atlas-based	Kwak et *al.* [53]	Hippocampus	MR T1
	Wu et *al.* [66]	Multi-structure	MR T1
Multiple Atlas-based	Aljabar et *al.* [67]	Multi-structure	MR T1
	Artaechevarria et *al.* [44]	Multi-structure	MR
	Asman et *al.* [69]	Multi-structure	MR
	Bondiau et *al.* [5]	Brainstem	MR T1,T2
	Cardoso et *al.* [52]	Hippocampus	MR T1
	Collins et *al.* [45]	Hippocampus, amygdala	MR T1
	Coupe et *al.* [46]	Multi-structure	MR T1
	Heckemann et *al.* [65]	Multi-structure	MR T1
	Khan et *al.* [49]	Hippocampus	MR T1
	Kim et *al.* [50]	Hippocampus	MR 7T
	Lotjonen et *al.* [68]	Multi-structure	MR T1
	Panda et *al.* [81]	Optic nerves, eye globes	CT
	Wang et *al.* [54]	Hippocampus	MR
	Zarpalas et *al.* [55]	Hippocampus	MR T1
Active Shape models	Bailleul et *al.* [61]	Multi-structure	MR
	Bernard et *al.* [112]	Subthalamic nucleus	MR T1
	Olveres et *al.* [58]	Mid Brain	MR T1, SWI
	Pitiot et *al.* [104]	Multi-structure	MR T1
	Rao et *al.* [107]	Multi-structure	MR
	Tu et *al.* [63]	Multi-structure	MR T1
	Zhao et *al.* [105]	Multi-structure	MR
Active Appearance models	Babalola et *al.* [62]	Multi-structure	MR T1
	Babalola et *al.* [109]	Multi-structure	MR T1
	Brejl et *al.* [101]	Corpus callosum, cerebellum	MR
	Cootes et *al.* [59]	Multi-structure	MR
	Duchesne et *al.* [60]	Medial temporal lobe	MR T1
	Hu et *al.* [48]	Hippocampus, amygdala	MR T1, T2
	Hu et *al.* [64]	Medial temporal lobe	MR T1
Parametric deformable models	Lee et *al.* [120]	Brainstem,cerebellum	MR
	Mcinerney et *al.* [119]	Corpus callosum,cerebellum	MR
	Mcintosh et *al.* [56]	Corpus callosum	MR
	Szekely et *al.* [70]	Multi-structure	MR
Geometric deformable models	Bekes et *al.* [124]	Eyeballs,lens,nerves	CT
	Duncan et *al.* [123]	Hippocampus	MR T1
	Ghanei et *al.* [40]	Hippocampus	MR
	Leventon et *al.* [57]	Corpus callosum	MR
	Shen et *al.* [41]	Hippocampus	MR T1
	Tsai et *al.* [71]	Multi-structure	MR
	Wang et *al.* [122]	Multi-structure	MR
	Yang et *al.* [72]	Multi-structure	MR
	Zhao et *al.* [51]	Hippocampus	MR
Machine Learning. ANN	Hult et *al.* [42]	Hippocampus	MR T1,T2
	Magnotta et *al.* [73]	Multi-structure	MR T1,T2
	Moghaddam et *al.* [128]	Putamen,caudate, thalamus	MR T1
	Pierson et *al.* [74]	Cerebellar subregions	MR T1,T2
	Powell et *al.* [76]	Multi-structure	MR T1,T2,PD
	Spinks et *al.* [126]	Thalamus,mediodorsal nucleus	MR T1,T2,PD
Machine Learning. SVM	Golland et *al.* [75]	Hippocampus,amygdala,corpus callosum	MR
	Morra et *al.* [43]	Hippocampus	MR T1
	Morra et *al.* [47]	Multi-structure	MR T1
	Powell et *al.* [76]	Multi-structure	MR T1,T2,PD

Table 3.1: Summary of subcortical structures segmentation methods.

Ref	Dataset	Training/ Testing	Reference contours	Image matrix (x y z)	Image resolution (x y z)	Others
Aljabar et *al.* [67]	275 subjects	LOOCV	n.a.	-×-×-	-×-×- mm^3	37.9 ± 21.1 y.o. [4-83]
Asman et *al.* [69]	15 subjects (OASIS[1])	-	n.a.	-×-×-	-×-×- mm^3	-
Babalola et *al.* [62]	270 subjects	27 LOOCV with 260 training and 10 testing	n.a.	-×-×- -×-×-	-×-×- mm^3	4.5-83 y.o.
Bernard et *al.* [112]	31 subjects	7 training 24 testing	4 experts	-×-×-	0.4688×0.4688×1.2 mm^3	-
Bondiau et *al.* [5]	26 subjects	20 training 6 testing	7 experts	T1 256×256×60 T2 256×256×64	T1 1×1×2 mm^3 T2 1×1×1.9 mm^3	-
Cardoso et *al.* [52]	55 subjects 30 subjects (ADNI[2])	LOOCV	1 n.a.	-×-×- -×-×-	-×-× -mm^3 -×-×1.2 mm^3	70 y.o. as average
Collins et *al.* [45]	152 subjects (ICBM[3])	80 (testing)	4 experts 5 times each	-×-×140	-×-×1 mm^3	25.09 ± 4.9 y.o. [18-42] 39 male/41 female
Cootes et *al.* [59]	28 subjects	LOOCV	1 expert	-×-×-	1×1×1.5 mm^3	n.a.
Coupe et *al.* [46]	80 subjects (ICBM[3])	LOOCV	1 expert	-×-×-	-×-×- mm^3	25.09 ± 4.9 y.o. 39 male/41 female
Duchesne et *al.* [60]	80 subjects (ICBM[3])	70 training 10 testing	n.a.	-×-×-	-×-×- mm^3	-
Duncan et *al.* [123]	12 subjects	LOOCV	n.a.	-×-×-	1.2×1.2×1.2 mm^3	14-43 y.o.
Ghanei et *al.* [40]	6 sets of 2D images	n.a.	1	-×-×-	-×-×- mm^3	-
Golland et *al.* [75]	30 subjects	LOOCV	Several experts	-×-×-	-×-×- mm^3	-
Heckemann et *al.* [65]	30 subjects	LOOCV	n.a.	192×256×124	-×-×-1.5 mm^3	30.5 y.o. [20-54] 15 male/15 female
Hu et *al.* [48]	80 subjects (ICBM[3])	60 training 20 testing	Several experts	-×-×-	T1 1×1×1 mm^3 T2 2×1×1 mm^3	18-35 y.o.
Hu et *al.* [64]	54 subjects (ICBM[3])	40 training 1 testing	Several experts	-×-×-	T1 1×1×1 mm^3	18-35 y.o.
Khan et *al.* [49]	69 subjects 37 elderly subjects	30 (training)	n.a. n.a.	256×256×160	-×-×-1 mm^3	40-44 y.o. (39 male/30 female) 70-84 y.o. (17 male/20 female)
Kim et *al.* [50]	6 subjects	LOOCV	n.a.	572×512×60	-×-×- mm^3	-
Kwak et *al.* [53]	27 Healthy subjects	-	1 expert	480×480×360	-×-×0.5 mm^3	71.82 ± 7.34 y.o.
Leventon et *al.* [57]	49 (2D) images	n.a.	n.a.	-×-×-	-×-×- mm^3	-

[1] Open Access Series of Imaging Studies
[2] Alzheimer's Disease Neuroimaging Initiative
[3] International Consortium for Brain Mapping
[4] Internet Brain Segmentation Repository
[5] Baltimore Longitudinal Study of Aging

Table 3.2: Summary of experimental set up of referenced segmentation works. Part I

Ref	Dataset	Training/ Testing	Reference contours	Image matrix (x y z)	Image resolution (x y z)	Others
Lotjonen et *al.* [68]	18 subjects (IBSR^4) 60 subjects	LOOCV 30 (Testing)	n.a. n.a	256×256×128 192×192×160 256×256×180	0.8-1×0.8-1×1.5 mm^3 0.9×0.9×1.2 mm^3 1.3×1.3×1.2 mm^3	7-71 y.o.
Magnotta et *al.* [73]	30 subjects	20 (Training) 10 (Testing)	2	T1 256×192×124 T2 256×192×-	T1 -×-×- mm^3 T2 -×-×3-4 mm^3	
Moghaddam et *al.* [128]	15 subjects (IBSR^4)	12 (Training) 3 (Testing	n.a.	-×-×-	-×-×- mm^3	-
Morra et *al.* [43]	120 subjects	27 (Training) 83 (Testing	n.a.	256×192×-	-×-×1.5 mm^3	-
Morra et *al.* [47]	70 subjects	30 (Training) 40 (Testing	n.a.	256×192×-	-×-×1.5 mm^3	-
Olveres et *al.* [58]	10 subjects	LOOCV	1 expert	-×-×-	-×-×- mm^3	-
Panda et *al.* [81]	543 images from 181 thyroid eye patients	30 independent for testing	2 experts	-×-×-	-×-×0.4-0.5 mm^3	49 y.o. [9-83] 81 % females
Pierson et *al.* [74]	30 subjects	20 (Training) 10 (Testing)	2	T1 256×192×124 T2 256×192×-	T1 -×-×- mm^3 T2 -×-×3-4 mm^3	
Pitiot et *al.* [104]	20 subjects	n.a.	Several experts	256×256×124	1×1×1 mm^3	-
Powell et *al.* [76]	25 subjects 15 subjects	15 train/10 test 10 train/5 test	Several experts	T1 256×256×124 T2 256×256×124 T1 256×192×124 T2 256×192×-	T1 -×-×- mm^3 T2 -×-×1.8 mm^3 T1 -×-×- mm^3 T2 -×-×3-4 mm^3	Evaluates multiple struct. Evaluate hippocampus
Rao et *al.* [107]	178 subjects	LOOCV	n.a.	-×-×-	1×1.5×1 mm^3	-
Shen et *al.* [41]	10 subjects (BLSA^5)	n.a.	2	256×256×124	0.9375×0.9375×1.5 mm^3	55-85 y.o.
Spinks et *al.* [126]	30 subjects	20 (Training) 20 (Testing)	n.a.	T1 256×192×- T1 256×192×-	T1 -×-×1.5 mm^3 T2 -×-×3 mm^3	15 males (28.33±8.46 y.o.) 15 females (28.00±8.21 y.o.)
Tu et *al.* [63]	15 subjects 10 subjects	8 train/7 test 5 train/5 test	n.a.	T1 256×256×124 T2 512×512×112	-×-×- mm^3	-
Wang et *al.* [122]	35 subjects (OASIS^1)	15 atlas and 20 testing	n.a.	-×-×-	-×-×- mm^3	n.a.
Wang et *al.* [54]	2012 MICCAI challenge brain images	n.a.	n.a.	-×-×-	-×-×- mm^3	n.a.
Wu et *al.* [66]	9 subjects (6 male/3 female) 13 subjects (IBSR^4)	LOOCV LOOCV	2 experts -	256×192×- 256×256×128	-×-×1.5 mm^3 -×-×1.5 mm^3	24.3 y.o. [20-32] -
Yang et *al.* [72]	12 subjects	LOOCV	n.a.	172×148×124	-×-×- mm^3	-

[1] Open Access Series of Imaging Studies
[2] Alzheimer's Disease Neuroimaging Initiative
[3] International Consortium for Brain Mapping
[4] Internet Brain Segmentation Repository
[5] Baltimore Longitudinal Study of Aging

Table 3.3: Summary of experimental set up of referenced segmentation works. Part II

Method	Benefits	Assumptions and/or Limitations
Single atlas-based	- Fast - Sufficient fot intra-patient segmentation	- Lower accuracy if there is significant anatomical variation
Multiple atlas-based	- Capable to cover a higher variability than with a single atlas - Combination of propagated labels may overcome limitations of single atlases - Atlases are easy to build	- Computationally expensive - Rely on the registration - Success also depends on atlas building
Active shape models	- Relatively fast - Easy to implement - Larger capture range than AAM - Robust against noise	- Cannot create unseen shapes - Not robust when different images are introduced - May not converge to a good solution
Active appearance models	- More powerful than ASM in detecting the global minima - Better match to image texture than ASM - Robust against noise	- Excessive usage of memory - Hard to implement - Cannot generalize well to unsampled population
Parametric deformable models	- No training required - Provide flexibility	- Sensitive to initialization - Susceptible to noise and artifacts
Geometric deformable models	- No training required - Provide flexibility - Ability to handle topological changes - Easily deform to highly complex structures	- Sensitive to initialization - Stopping criteria hard to define - May get stuck in any local minima
Artificial neural networks	- can be used for classification or regression - able to represent Boolean functions - tolerant of noisy inputs - instances can be classified by more than one output	- difficult to understand structure of the algorithm - too many attributes can result in overfitting - optimal network structure can only be determined by experimentation
Support vector machines	- models nonlinear class boundaries - overfitting is unlikely to occur - computational complexity reduced to quadratic optimization problem - easy to control complexity of decision rule and frequency of error	- training is slow compared to other ML approaches - difficult to determine optimal parameters when training data is not linearly separable - difficult to understand structure of the algorithm

Table 3.4: Summary of benefits, assumptions and limitations of different segmentation methods for brain structures.

Chapter 4
Our Contribution

“ I have not failed. I have just found 10.000 ways that will not work.”
Thomas A. Edison

This chapter introduces the main contributions of this thesis. Typical setting of a machine learning classifier mainly involves two elements: the learning method and the set of features. On the one hand, we propose to employ a stack of denoised auto-encoders in a deep fashion to segment the OARs. On the other hand, we propose the use of new features to achieve better performance. These two components of the classifier are the cornerstone of our dissertation. To understand the context of our proposal, some fundamental notions of machine learning, such as the representation of the data and the classification task, are briefly presented in the first section,. Next, the deep learning technique employed in our work is introduced. Following, in section 4.3, features proposed throughout our work are detailed. And in the last section of this chapter the steps to train the deep network are explained.

4.1 Introduction to Machine Learning

The endeavor to understand intelligence implies building theories and models of brains and minds, both natural as well as artificial. From the earliest writings of India and Greece, this has been a central problem in philosophy. With the arrival of the digital computer in the 1950’s, this became a central concern of computer scientists as well. Thanks to the parallel development of the theory of computation, a new set of tools with which to approach the problem through analysis, design, and evaluation of computers and programs exhibiting some aspects of intelligent behavior was provided. The ability to recognize and classify patterns or to learn from experience were some of these intelligent behaviors [144].

Among the different ways to define the notion of intelligence, we interpret it as the ability to take the right decisions, according to some criterion. Taking appropriate decisions generally requires some sort of knowledge that is utilized to interpret sensory data. Decisions are then taken based on that information. Nowadays, as a result of all the programs that humans have crafted, computers possess somehow their own intelligence. This understanding allows computers to easily carry out tasks that might be intellectually difficult for human beings. Nevertheless, tasks that are effortlessly done by humans and animals might still remain unreachable for computers. Many of these tasks fall under the label of Artificial Intelligence (AI).

Reasons of failure in such tasks can be summarized in the lack of explicit information when trying to transfer the knowledge to the machine. That is, in other words, in situations where to solve a given problem a computer program cannot be directly written. This commonly occurs when we, humans, know how to perform an action or a task and are not able to explain our expertise. Learning is therefore required by the machine to execute such a task. In this way, computers learn from experience and understand the world in terms of a hierarchy of concepts, where each concept is defined in terms of its relation to simpler concepts. This hierarchical distribution will allow the computer to learn complex concepts by depicting them with simpler ones. The capability of AI-based systems to acquire their own knowledge by extracting patterns from raw data is known as *machine learning* (ML). Consider as example the problem of speech recognition. This task can be done apparently without any difficulty, but explanation on how we do it is not straightforward. Due to differences in gender, age or accent, for example, there exists a speaker variability, which makes different people utter the same word differently. We can easily recognize who speaks, or to which kind of population a given utterance belongs because of our experience. Nevertheless, in machine learning, the approach consists on collecting a large collection of sample utterances from different people and learning how to map all these to words. Thus, the machine learns how to automatically extract the algorithm to perform this task. In short, machine learning involves training a computer system to perform some task, rather than directly programming the system to perform the task.

4.1.1 The "Task"

One of the main strengths for which machine learning has been especially interesting is the variety of tasks that can be achieved with it. From an engineering point of view, ML has brought us the capability to approach some tasks that would be too hard to solve with hand-crafted computer programs. On the other hand, from a scientific point of view, understanding machine learning has provided us the knowledge of the principles that govern intelligent behavior, which establishes the basis to accomplish certain tasks.

Hence, the learning process itself is not the aforementioned task. Learning is the process of obtaining the ability to achieve the task. For instance, if we want a car to be able to autonomously drive, then driving is the task. To complete the task we could either program the car to learn to drive, or directly write a computer program that specifies how manually drive, instead. We therefore understand that machine learning can be employed to solve many kinds of tasks. Nevertheless, one of the most common tasks, which is also the task to perform in this dissertation, is classification.

Classification entails assigning an observation to a category or class. To solve this task, the learning algorithm is typically asked to build a function $f\colon \mathbb{R}^n \to \{1, ..., k\}$ which can be applied to any input. The output of this function, $f(x)$, can be then interpreted as an estimation of the class to which x belongs to. Methods used for classification often predict the probability of an observation of each of the categories, or classes, of a qualitative variable as the basis for later on providing the classification. Let's consider object recognition as example of classification. Here, an image usually represents the input, x, and the output, $f(x)$, is a numeric value which identifies the object in the image.

Learning how to classify objects to one of a pre-specified set of categories or classes is a characteristic of intelligence that has been of keen interest to researchers in psychology and computer science. Identifying the common core characteristics of a set of objects that are representative of their class is of enormous use in focusing the attention of a person or computer program. For example, to determine whether an animal is a zebra, people know that looking for stripes is much more meaningful rather than examining its tail or ears. Stripes alone are not sufficient to form a class description for zebras, since tigers have them also, but they are certainly one of the important characteristics. Thus, stripes strongly figure in our concept or generalization of what zebras are. The ability to perform classification and to be able to learn to classify gives people and computer programs the power to make decisions. The efficacy of these decisions is affected by performance on the classification task, which in turn strongly depends on the representation of the data.

4.1.2 Data Representation

The choice of data representation plays a crucial role on the performance of a ML-based classifier. In a typical machine learning task, data is represented as a table of examples or instances. Each instance is described by a fixed number of measurements, or features, along with a label that denotes its class. Features, which are also sometimes called attributes, are typically one of two types: nominal or numeric. While the former are members of an unordered set, the later are represented by real numbers. Table 4.1 shows ten instances of benign and malignant tumors according to some of their characteristics. Each instance is a tumor described in terms of the attributes *size*, *homogeneity* and *shape*, along with the class label which indicates whether a tumor is benign or malignant. During learning, correlation between these features and various outcomes will be learned, and this will be employed to make predictions on new unseen instances.

#instances	Features			Tumor Type
	Size	Homogeneity	Shape	
1	Small	Yes	Circular	Benign
2	Medium	Yes	Irregular	Malignant
3	Medium	No	Irregular	Malignant
4	Large	Yes	Circular	Benign
5	Small	No	Irregular	Malignant
6	Large	No	Irregular	Malignant
7	Medium	No	Circular	Malignant
8	Medium	Yes	Circular	Benign
9	Small	Yes	Irregular	Benign
10	Small	No	Circular	Malignant

Table 4.1: Brain tumor classification table. Some tumor properties are used as features to train the classifier.

To illustrate the importance of selecting the proper representation of the data for a given problem, two different representations of the same data are shown in figure 4.1. Available data is sampled according to points location or coordinates. A simple classification task would be to separate the two data categories by just drawing a line between the two groups. However, whilst on the example where data is represented by Cartesian coordinates the task is impossible, in the example representing the data with polar coordinates the task becomes simple to solve with a vertical line.

This dependence on data representations is a phenomenon that commonly appears throughout computer science. Operations such as searching a collection of data can proceed exponentially faster if the collection is structured and indexed intelligently. Thus, we can assume that many AI tasks can be easily solved by designing the proper set of features for a specific task. For example, as illustrated in the case of tumor characterization (table 4.1), a useful feature for representing a tumor is its shape. It may be useful for tumor characterization because type of tumor it is often, together with other factors, determined by the nature of its shape. Shape gives therefore a strong clue as to whether a tumor is benign or malign.

However, for many tasks, knowing which features should be extracted is not trivial. For instance, following the tumor example, suppose that we would

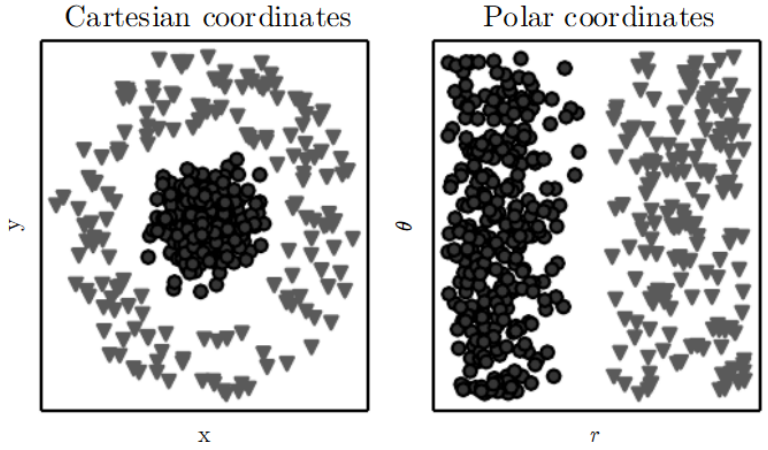


Figure 4.1: Example of different representations: a) Cartesian coordinates and b) polar coordinates are used to represent the data.

like to write a computer program to detect tumors in medical images. We, or doctors, know how tumors may look like. So we might like to use the appearance of a tumor as a feature. Unfortunately, it is very difficult to exactly describe how a tumor looks like in terms of pixel values. This is particularly harder when combining multiple image sequences. One solution to tackle this problem is to use ML to discover not only the mapping from representation to an output, but also the data representation itself. This approach is known as *representation learning*.

We have seen that, in general, a good data representation is the one that makes the further learning task easier. Generally, designing features aims at separating the factors of variation that explain the observed data. Hence, hand-designed representations of the data usually provide satisfactory classification performances. Nevertheless, learned representations typically result in much better performance, since the best data configuration is represented in a more compressed and meaningful way. Despite there exist sophisticated algorithms to learn data representations, factors or sources of variation still introduce a major source of difficulty in many real world AI applications: they influence every single piece of observed data. In such situations, factors of variations must be *unscrambled* and careless factors discarded. Nevertheless, this is not straightforward and complex understanding of the data is required to identify such high-level abstract features. To solve this main issue in representation learning, representations that are expressed in terms of other, simpler representations are introduced. And this is exploited in *deep learning*, which will be detailed later on.

4.1.3 Learning Algorithms

A learning algorithm, or an induction algorithm, forms concept descriptions from known data or experience. Concept descriptions are often referred to as the knowledge or model that the learning algorithm has induced from the input data. It models then the function that will perform the classification task from the representation of the given data. Knowledge may be represented differently from one algorithm to another.

Advantageously, while most conventional computer programs are explicitly programmed for each process, ML-based systems are able to learn a given task, regardless of its complexity. By following the lemma "divide and conquer", a complex problem can be decomposed into simpler tasks, in order to be able to understand it and solve it. Artificial Neural Networks (ANN), represent one approach to achieve this. An ANN is a massively parallel computing system consisting of an extremely large number of simple processors, i.e. neurons, with many interconnections between them, i.e. weights. Learning in ANN is performed by using algorithms designed to optimize the strength of the connections in the networks. A network can be subject to supervised or unsupervised learning. In order to be referred to as supervised learning, an external criteria has to be used and matched by the network output. Otherwise, learning is termed as unsupervised, or also self-organizing. In this approach, no sample outputs are provided to the network against which it can measure its predictive performance of a given vector of inputs. As a result, there exist more interaction between neurons. Interaction is often performed by employing feedback and intralayer connections between neurons, which promotes self-organization. A detailed explanation of ANNs them can be found in Appendix A.

The purpose of this section is to review the deep learning technique explored, which can be applied to the problem of segmenting critical structures on MRI scans during the radiation treatment planning for brain cancer. Throughout this thesis, two learning algorithms are used as a basis for comparison between their performance. The first of these learning algorithms is Support Vector Machines, which constitutes one of the most successful classifiers inside the classic machine learning techniques. Nevertheless, it is important to note that during the research conducted for this work, it has been found that there is a gap missing in the state-of-the-art, as no deep architectures seem to have been fully explored yet to tackle the problem of brain structures segmentation on MRI. We try to make a step towards this direction in the framework of this project, and we propose the use of a deep learning classification system based on Stacked Denoising Autoencoders (SDAE). Since SVM does not represent the core of this thesis, and it is only employed for comparison purposes, a theoretical introduction has been included in Appendix B.

4.2 Deep Learning

Deep Learning is a new subfield of machine learning that focuses on learning deep hierarchical models of data. Modern deep learning research takes a lot of its inspiration from neural network research of previous decades. Whereas most current learning algorithms correspond to shallow architectures with 1 up to 3 levels of abstractions, the mammal brain is organized in a deep architecture with multiple levels, each level corresponding to a different cortex region. Inspired by the architectural depth of the brain, neural network researchers had wanted for decades to train deep multilayer neural networks, but no successful attempts were reported before 2006 (except convolutional NNs).

4.2.1 Historical context

Inspired by the understanding of biological neurons, straightforward algorithms were proposed to create artificial neural networks in the 60's [145]. Although this discovery created a great excitement and expectations over the scientific community, initial enthusiasm soon declived because of the inability of these simple learning algorithms to learn representations. This shortcoming in learning what the hidden layers of the network should represent led to a strong influence of symbolic computation and expert systems in the Artificial Intelligence domain during the subsequent years. The introduction of the backpropagation algorithm to learn patterns that were not linearly-separable [146] made possible the use neural networks to solve problems which were previously insoluble. This caused a replenishment on the research of neural networks. Lately, in the 90's and 2000's, and despite the remarkable results of artificial neural networks to perform some tasks [147], some other approaches dominated the field [141, 143, 148].

One of the main reasons to abandon artificial neural networks in favor of these more limited approaches was the difficulty of training deep networks. Training deep architectures was a difficult task and classical methods that had proved to be effective when applied to shallow architectures were not as efficient when adapted to deep architectures. Simply adding more layers did not necessarily lead to better solutions. On the contrary, as the number of hidden layers increased -i.e. architecture got deeper- it become more difficult to obtain good generalization. For example, the deeper the network, the lesser the impact of the back-propagation algorithm on the first layers. The error from the output layer that was back propagated to the inner layer was getting smaller at each time a layer was passed over, making that the multilayer network in fact did not learn. Gradient-based training of deep supervised multi-layer neural networks starting from random initialization then tended to get stuck in local minima [149]. Additionally, a neural network composed by three layers -i.e. only a hidden layer- was mathematically demonstrated to be a universal approximator [150]. As a consequence, solutions obtained with deeper networks corresponded to poor solutions, with worse performance than shallow networks. Hence, until some years ago, most machine learning techniques exploited shallow structures architectures, where networks were typically limited to one or two hidden layers.

However, it was not until 2006 when the concept of *Greedy Layer-Wise Learning* was introduced [149, 151, 152]. This new concept profits from a semi-unsupervised learning procedure. Unsupervised learning is used in a first stage to initialize the parameters of the layers, one layer at a time, and then a fine-tuning of the whole system is done by a supervised task. Since then, deep structured learning, or more commonly known as deep learning or hierarchical learning, has emerged as a new area of machine learning research [151, 153], impacting a wide range of research fields.

4.2.2 Advantages respect to shallow architectures

We have seen that a simple neural network with two hidden layers already theoretically represents a universal function approximator capable of approximating any function to any arbitrary accuracy. However, one of the main benefits of using deep networks comes from the side of computational efficiency. Indeed, complex functions can often be approximated with the same accuracy using a deeper network that has much fewer total number of units compared to a typical two-hidden-layer network containing large hidden layers. The size of the training set is often a limiting factor when using neural networks based systems. By employing deeper network instead, models with smaller degree of freedom, which require smaller datasets to train [154], are built. This leads to a shrinkage on the training dataset size required.

Another, probably more compelling, factor is that typical approaches for classification must be generally preceded by a feature selection step, where most discriminative features are privileged for a given problem. Such step, however, is not needed in deep learning-based classification schemes. What differentiates deep learning approaches from other conventional machine learning techniques, therefore, is their ability to automatically learn features from data which largely contributes to improvements in terms of accuracy. In other words, deep learning learns a better and more compact representation of the input data. This represents an important advantage and removes a level of subjectivity from conventional approaches, where the researcher typically has to decide which set of features must be tried. With the inclusion of deep learning techniques in the classification scheme this step is thus avoided.

Furthermore, as it has been shown in the previous section, one of the problems of classical shallow networks is its difficulty to train networks with more than two or three hidden layers. By employing a learning algorithm that greedily trains one layer at time deeper networks can be used. Apart from allowing the use of networks with more hidden layers, pre-training each layer with an unsupervised learning algorithm might result in the achievement of much better results [155, 156]. Unsupervised pre-training allows, indeed, to achieve good generalization performance when the training set is limited in terms of size by positioning the network in a region of the parameter space where the supervised gradient descent is less likely to drop in a local minimum of the loss function.

A worthy point to highlight is that deep learning approaches are recently breaking records in several domains, such as speech, signal, image and text mining and recognition and improving state of the art classification methods in accuracy by, sometimes, more than 30 %, where the prior decade struggled to barely achieve 1-2 % of improvements [157, 158].

The main shortcoming of deep learning techniques, which is actually one of its advantages, is the large amount of data required to unsupervisedly craft the features during its first stage.

4.2.3 Different levels of abstraction

Following the analogy with the human brain, the process of object recognition in the visual cortex begins in the low-level primary area V1. Then, the process proceeds in a roughly bottom-up fashion through areas V2 and V4, ending in the inferotemporal cortex (IT), figure 4.2. Once the information reaches the IT, it travels to prefrontal areas, where it plays a role in perception, action, planning and memory. These hierarchically organized circuits in the human brain exploit circuit modularity and reuse general subcircuits in order to economize on space and energy consumption. Thus, in a hierarchical model, lower layers might include dictionaries of features that are general and yet applicable in the context of many specific classification tasks.

We have seen that deep learning is a kind of representation learning in which there are multiple levels of features. These features are automatically discovered and they are composed together in the various levels to produce the output. Each level represents abstract features that are discovered from

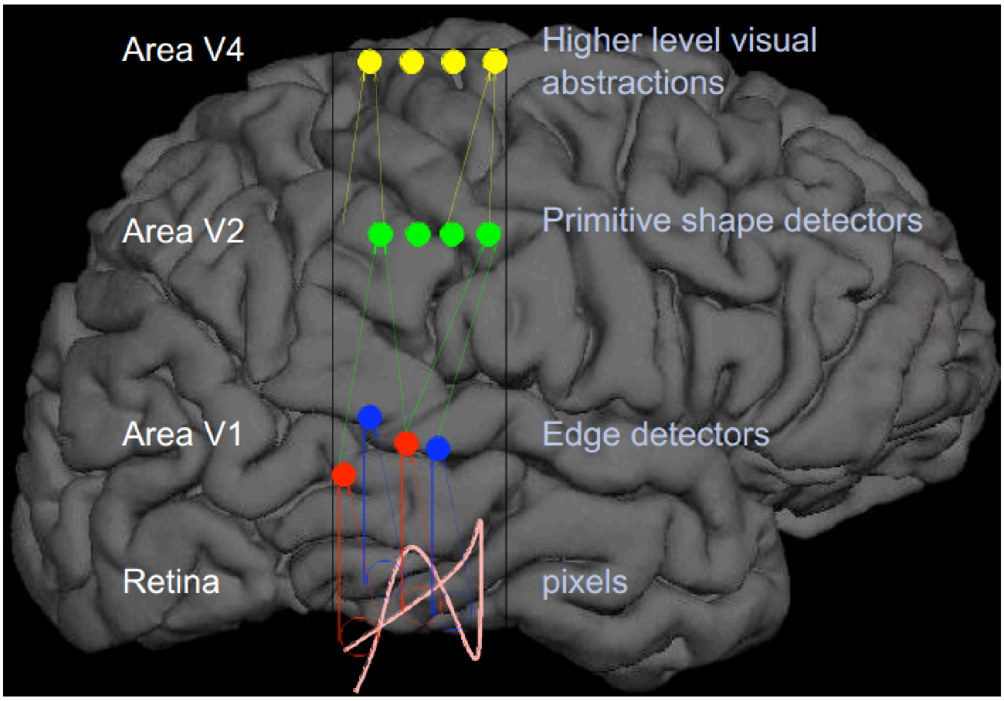


Figure 4.2: Deep architecture of the brain.

the features represented in the previous level. Hence, the level of abstraction increases with each level. This type of learning enables discovering and representing higher-level abstractions. In neural networks, the multiple layers correspond to multiple levels of features. These multiple layers compose the features to produce the output. While the first layers use to be more generic, last layers are often strongly task-specific. Therefore, the higher the layer, the more specialized the features are.

4.2.4 Convolutional neural networks

Among all the deep learning approaches, convolutional neural networks (CNNs) have demonstrated to be very powerful when classifying medical images. These artificial networks are made up of convolutional, pooling and fully-connected layers. These type of networks are mainly characterized by three main properties: local connectivity of the hidden units, parameter sharing and the use of pooling operations.

A CNN consists of a succession of layers which perform several operations on the input data. First, convolutional layers C convolve images presented at their inputs with a predefined number of kernels, k. These kernels have a certain size, s, and are typically followed by activation units that rescale the convolution results in a non-linear manner. Pooling layers reduce the dimensionality of the responses produced by the convolutional layers through downsampling. Different strategies can be adopted to perform the pooling: average or max-pooling, for example. At the end, fully connected layers are the responsible of extracting compact, high level features from the data. A typical workflow for a convolutional neural network is shown in figure 4.3. In these networks, two or three-dimensional patches are commonly fed into the deep network, which unsupervisedly learns the best features representation of those given patches. In other words, it learns a hierarchical representation of the input data and is able to decode the important information contained on the data. By doing this, a deep network is able to provide a hierarchical feature representation of each patch and ensure discriminative power for the learned features. Networks based on convolutional filters, i.e. CNN, perfectly suit to deal with data presenting a grid structured representation, such as 2D or 3D image patches. However, when input data composed by features not presenting a grid-based representation is employed, CNNs might not represent the best solution.

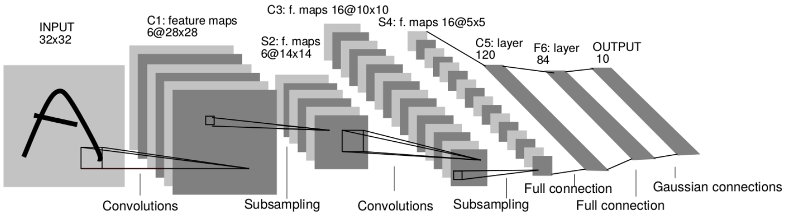


Figure 4.3: A typical workflow for a convolutional neural network.

Valuable information inherited from classical machine learning approaches to segment brain structures is not therefore included into the CNNs. This knowledge may come in the form of likelihood voxel values, voxel location, as well as textural information, for example, which is greatly useful to segment structures that share similar intensity properties. Because we wish to employ arrays composed by concatenation of different features, which will be introduced in Section 4.3, we consider the use of denoised auto encoders (DAE) instead, which is able to deal with such type of features arrays. Another reason for employing DAE is because of the limited size of the number of training and labeled data. Instead of random initialization of the network weights, values are obtained by using DAEs which act as a pre-training step in an unsupervised fashion. Thanks to this the network can be trained with such limited amount of data while avoiding overfitting.

4.2.5 Auto-Encoders

Autoencoders are a method for performing representation learning, an unsupervised pretraining process during which a more useful representation of the input data is automatically determined. Representation learning is important in machine learning since the performance of machine learning methods is heavily dependent on the choice of data representation in which they are applied. For many supervised classification tasks, the high dimensionality of the input data means that the classifier requires a huge number of training examples in order to generalize well and not overfit. One solution is to use unsupervised pretraining to learn a good representation for the input data and during actual training, transform the input examples into an easier form for the classifier to learn. Autoencoders are one such representation learning tool.

Classical auto-encoders (AE) have been recently developed in the deep learning literature in different forms [159]. In its simplest representation, an AE is formed by two components: an encoder $h(\cdot)$ that maps the input $x \in R_d$ to some hidden representation $h(x) \in R_dh$, and a decoder $g(\cdot)$, which maps the hidden representation back to a reconstructed version of the input x, so that $g(h(x)) \approx x$ (Fig. 4.4). Therefore, an AE is trained to minimize the discrepancy between the data and its reconstruction. This discrepancy represents the difference between the actual output vector and the expected output vector that is the same as the input vector. As a result, AEs offer a method to automatically learn features from unlabeled data, allowing for unsupervised learning.

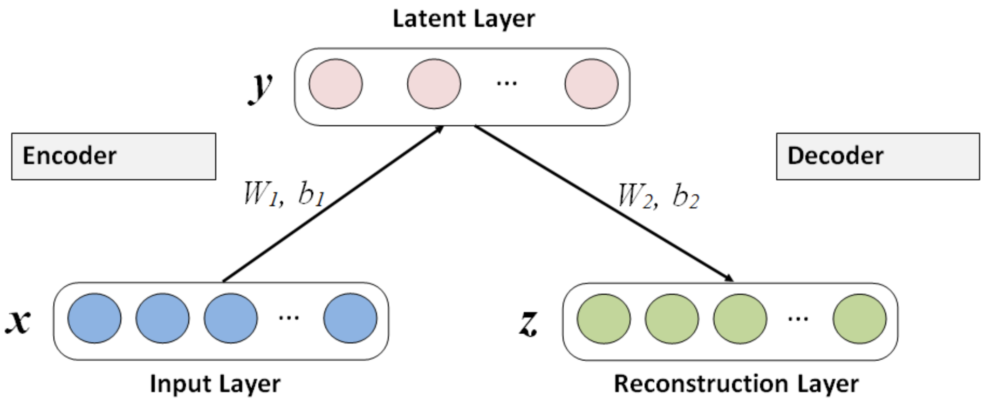


Figure 4.4: Auto-Encoder.

Let formulate an autoencoder in more detail. When a traditional autoencoder takes an input $x \in [0,1]^d$, first thing that it does it to map this input

-with the encoder- to a hidden representation $y \in [0,1]^{d'}$ through a deterministic mapping, as follows

$$y = f_\theta(x) = s(Wx + b) \tag{4.1}$$

which is parameterized by $\theta = \{W, b\}$. In addition, s is a non-linearity function, such as the sigmoid, W is a $d' \times d$ weight matrix and b is the bias vector. The resulting latent representation y is then mapped back -with the decoder- to a "reconstructed" vector $z \in [0,1]^d$ of the same shape as x. This reconstruction is defined as

$$z = g_{\theta'}(y) = s(W'y + b') \tag{4.2}$$

where parameterization is given by $\theta' = \{W', b'\}$ in this case. The weight matrix W' of the reverse mapping may optionally be constrained by $W' = W^T$ to be the transpose of the forward mapping. If this happens, the auto-encoder is said to have *tied weights*. Each training $x^{(i)}$ is thus mapped to a corresponding $y^{(i)}$ and a reconstruction $z^{(i)}$. In other words, z can be seen as a prediction of the input x, given the latent representation y. The parameters of this model are optimized such that the average reconstruction error is minimized

$$\begin{aligned} \theta', \theta'^{*} &= \arg\min_{\theta', \theta'^{*}} \frac{1}{n} \sum_{i=1}^{n} L(x^{(i)}, z^{(i)}) \\ &= \arg\min_{\theta', \theta'^{*}} \frac{1}{n} \sum_{i=1}^{n} L(x^{(i)}, z^{(i)}) \end{aligned} \tag{4.3}$$

where $L(\cdot)$ is a loss function such as the traditional squared error (for real-valued x)

$$L(x, y) = \|x - z\|^2 \tag{4.4}$$

Alternative loss functions can be used in 4.3. For example, if x and z are interpreted as either bit of vectors of bit probabilities, the cross entropy loss reconstruction can be used

$$L_H(x, y) = -\sum_{k-1}^{d} [x_k \log z_k + (1 - x_k) \log(1 - z_k)] \tag{4.5}$$

Using the cross entropy reconstruction formulation in 4.3, the average reconstruction error can be defined then as

$$\theta', \theta'^{*} = \arg\min_{\theta', \theta'^{*}} \mathbb{E}_{q^0(X)}[L_H(X, g_{\theta'}(f_\theta(X)))] \tag{4.6}$$

where $q^0(X)$ denotes the empirical distribution associated to the n training inputs and $\mathbb{E}$ refers to the Expectation

$$\mathbb{E}_{p(X)}[f(X)] = \int p(x)f(x)dx \tag{4.7}$$

To compute the Expectation, Eq. 4.7, we have assumed X and Y to be two random variables with joint probability density $p(X, Y)$, with marginal distributions $p(X)$ and $p(Y)$. Note that in the general auto-encoder framework, other forms of parameterized functions for the encoder or decoder, as well as other suitable choices of the loss function (corresponding to a different $p(X, Y)$ may be used. In particular, the usefulness of a more complex encoding functions was investigated in [160]. According to all this, it can be said that training an auto-encoder to minimize reconstruction error amounts to maximize a lower bound on the mutual information between input X and learned representation Y. Intuitively, if a representation allows a good reconstruction of its input, it means that it has retained much of the information that was present in that input.

The autoencoder yields lower reconstruction errors than other related batch algorithms based on matrix factorization. It efficiently generalizes to new inputs very accurately, with no expensive computations. This makes autoencoders fundamentally different from classical matrix factorization techniques. An example of neural encoding of an input and its corresponding reconstruction is shown in figure 4.5. In this figure, a reconstruction example of a handwritten digit input by employing neural encoding is shown. The input is represented by x, while $\hat{x}$ symbolizes its reconstruction. The input and the output are connected with the hidden layer h by the weights W and W^T, respectively. Weights W are responsible of encoding the input through the hidden layer, whereas weights W^T will decode the information in the hidden layer through the output, or reconstructed input.

4.2.6 Denoising Auto-Encoders

One serious potential issue when working with AE is that if there is no other constraint besides minimizing the reconstruction error (4.3), then an AE with n inputs and an encoding of dimension at least n could potentially just learn the identity function, for which many encodings would be useless, leading to just copy the input. That means that an AE would not differentiate test examples from other input configurations. There are different ways that an AE with more hidden units than inputs could be prevented from learning the identity, and still capture some valuable information about the input in its hidden representation. Adding randomness in the transformation from input

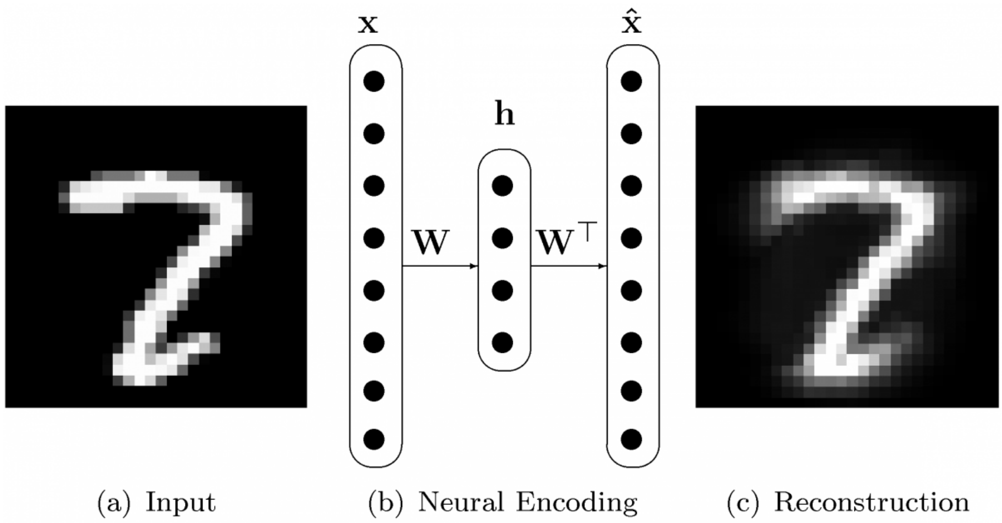


Figure 4.5: Reconstruction example of a handwritten digit input by employing neural encoding.

to reconstruction is one option, which is exploited in Denoising Auto-Encoders (DAEs) [161–166]. To force a hidden layer to discover more robust features and prevent it from simply learning the identity, a slight modification to the normal AE setup is done by corrupting the input x before mapping them into the hidden representation. This leads to a partially destroyed version $\tilde{x}$ by means of a stochastic mapping $\tilde{x} \sim q_D(\tilde{x}|x)$. Therefore, to convert an AE class into a DAE class, only adding a stochastic corruption step that modifies the input is required, which can be done in many ways.

Thus, following the formulation in classical AE in Section 4.2.5, the corrupted input $\tilde{x}$ is mapped to a hidden representation

$$y = f_\theta(x) = s(W_{\tilde{x}} + b) \tag{4.8}$$

from which z can be reconstructed (Figure 4.6).

$$z = g_{\theta'}(y) = s(W'_y + b') \tag{4.9}$$

As before, the parameters of the model are trained to minimize the average reconstruction error $L_H(x, z)$ (4.5) over a training set.

Hence the DAE tries to predict the corrupted values from the uncorrupted values, for randomly selected subsets of missing patterns, i.e., corrupted. The DAE is therefore a stochastic version of the AE.

Let define now the following joint distribution

$$q^0(X, \tilde{X}, Y) = q^0(X) q_D(\tilde{X} \| X) \delta_{f_\theta(\tilde{X})}(Y) \tag{4.10}$$

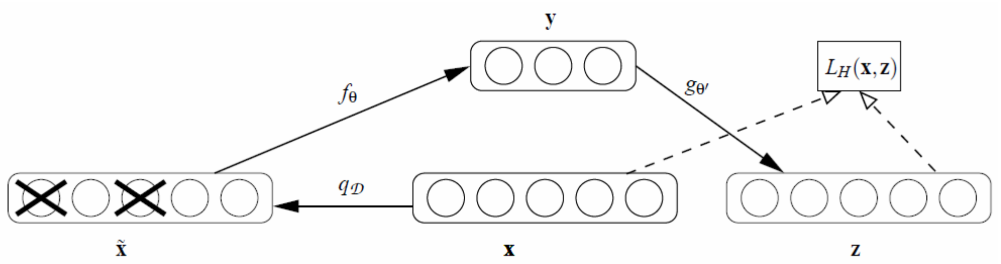


Figure 4.6: The denoising autoencoder architecture.

where $\delta_u(v)$ puts mass 0 when $u \neq v$. Thus Y is a deterministic function of $\tilde{X}$. Note also that the joint distribution function $q^0(X, \tilde{X}, Y)$ is parameterized by θ. The objective function minimized by the stochastic gradient descent becomes

$$\arg\min_{\theta',\theta'*} \mathbb{E}_{q^0(X,\tilde{X})}[L_H(X, g_{\theta'}(f_\theta(\tilde{X})))] \tag{4.11}$$

Therefore, from the point of view of the stochastic gradient descent algorithm, in addition to picking an input sample from the training set, we will also produce a random corrupted version of it, and take a gradient step towards reconstructing the uncorrupted version from the corrupted version. In this way, the denoising auto-encoder cannot learn the identity, unlike the basic auto-encoder, thus removing the constraint that $d' < d$ or the need to regularize specifically to avoid such a trivial solution.

Types of corruption. Corruption processes can be incorporated in many ways. The most common corruption processes are:

- Additive isotropic *Gaussian noise* (GS): $\tilde{x}||x \sim \mathcal{N}(x, \sigma^2 I)$;
- *Masking noise* (MN): a fraction v of the elements of the input x, that can be randomly selected, is forced to be 0;
- *Salt-and-pepper noise* (SP): a fraction v of the elements of the input x, that can be randomly selected, is set to their minimum or maximum possible value (typically 0 or 1) according to a fair coin flip.

Additive Gaussian noise is a very common noise model, and is a natural choice for real valued inputs. The *salt-and-pepper noise* will also be considered, as it is a natural choice for input domains which are interpretable as binary or near binary such as black and white images or the representations produced at the hidden layer after a sigmoid squashing function. For example, in [159], the stochastic corruption process consists in randomly setting some of the inputs to zero.

4.2.7 Stacked Denoising Auto-Encoders

Several DAEs can be stacked to form a deep network by feeding the hidden representation of the DAE found on the layer below as input to the current layer [159] (Figure 4.7), leading to what is known as Stacked Denoising Auto-encoder (SDAE). On this configuration, DAEs are stacked and trained bottom up in unsupervised fashion, followed by a supervised learning phase to train the top layer and fine-tune the entire architecture.

Weights between layers of the network are initially learned via an unsupervised pre-training step. Unsupervised pre-training of such architecture is done greedily, i.e. one layer at a time. Each layer is trained as a DAE by minimizing the reconstruction of its input. Once the first k layers are trained, the $(k+1)^{th}$ layer can be trained because the latent representation from the layer below can be then computed.

Once all the weights of the network are unsupervisedly computed, the highest level of the output network representation can be fed into a stand-alone supervised algorithm. Alternatively, and as in this work, a logistic regression layer can be added on top of the encoders. This yields a deep neural network amenable to supervised learning. Thus, the network goes through a second stage of training called *fine-tuning*, where prediction error is minimized on a supervised task [159]. A gradient-based procedure such as stochastic gradient descent is employed in this stage. The hope is that the unsupervised initialization in a greedy layer-wise fashion has put the parameters of all the layers in a region of parameter space from which a good local optimum can be reached by local descent. The unsupervised pre-training helps to mitigate the difficult optimization problem of deep networks by better initializing the weights of all layers [149].

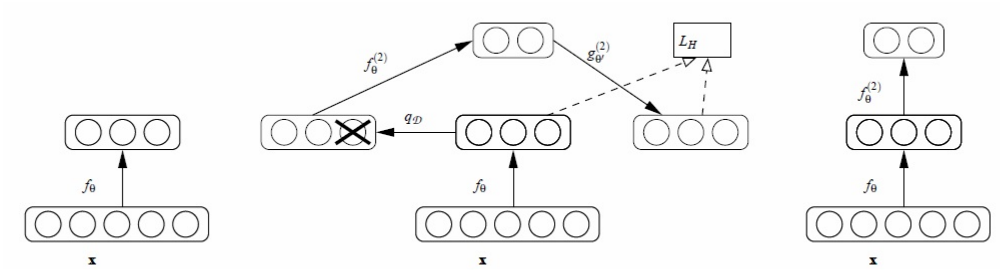


Figure 4.7: Stacked Denoising Auto-encoder. After training a first level denoising autoencoder (Fig. 4.6) its learnt encoding function f_θ is used on clean input (left). The resulting representation is used to train a second level denoising autoencoder (middle) to learn a second level encoding function $f_\theta^{(2)}$. From there, the procedure can be repeated (right) [159].

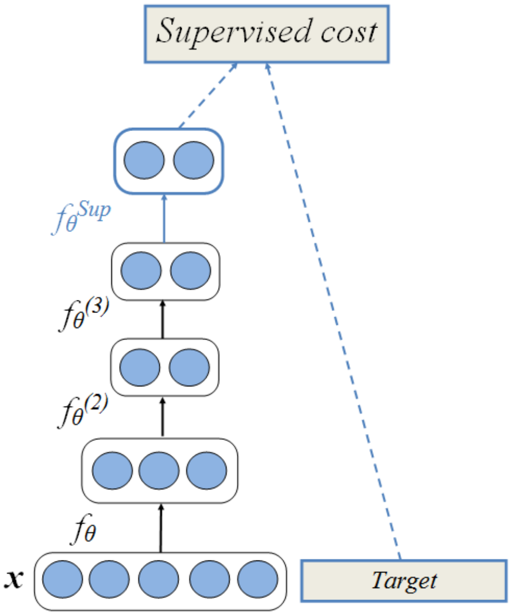


Figure 4.8: Fine-tuning of a deep network for classification. After training a stack of encoders as explained in the previous figure, an output layer is added on top of the stack. The parameters of the whole system are fine-tuned to minimize the error in predicting the supervised target (e.g., class), by performing gradient descent on a supervised cost [159].

4.2.7.1 Logistic Regression

Regression analysis is a field of mathematical statistics well explored which has been used for many years. In this type of analysis, given a set of observations, regression analysis can be employed to find a model that best fits the observation data. For instance, in linear regression, given an example i^{th} of a set of samples, x, a value for $y^{(i)}$ is predicted by a linear function $y = h_\theta(x) = \theta^\top x$. Although the linear regression model is simple and used frequently it is not adequate for some purposes, such as our goal, i.e. classification. Here, we aim at trying to predict binary values, such as labels ($y^{(i)} \in 0, 1$). A linear model has no bounds on what values the response variable can take, and hence y can take on arbitrary large or small values. However, it is desirable to bound the response to values between 0 and 1. For this we would need something more powerful than linear regression. In logistic regression a different hypothesis class used to predict the probability that a given sample belongs to the class A ('1') versus the probability that it belongs to the class B ('0') is employed. Particularly, the function learned will be of the form:

$$\begin{aligned} P(y=1 \mid x) = h_\theta(x) = \frac{1}{1+exp(-\theta^\top x)} \equiv \sigma(\theta^\top x) \\ P(y=0 \mid x) = 1 - P(y=1 \mid x) = 1 - h_\theta(x) \end{aligned} \tag{4.12}$$

The function $\sigma(z) \equiv \dfrac{1}{1+exp(-z)}$ is widely employed, and often referred to as sigmoid or logistic function. This function squeezes the value of $h_\theta(x)$ into the range [0,1]. By doing that, $h_\theta(x)$ can be interpreted as a probability. The goal is therefore to search a value of θ that makes the probability $P(y=1 \mid x)$ large when x belongs to the class A and small if x belongs to the class B instead. Imagine that we have a set of training samples with binary labels, $(x(i), y(i)) : i = 1, ..., m)$. To measure how well a given hypothesis $h_\theta(x)$ fits the training dataset, a cost function is defined as follows:

$$J(\theta) = -\sum_i \left(y^{(i)} \log(h_\theta(x^{(i)})) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))\right) \tag{4.13}$$

The next step is to learn to classify our training data by minimizing $J(\theta)$ in order to find the best choice of θ. Once training has been performed, new points can be classified either as class A or B by simply checking which of these classes is most probable. Basically, if for a given sample $P(y=1 \mid x) > P(y=0 \mid x)$, it will be labeled as class A('1'). Otherwise, it will belong to class B('0'). To minimize $J(\theta)$ the same tools typically employed for linear regression can be applied. This means to provide a function that computes $J(\theta)$ and $\nabla\theta J(\theta)$ for any request of the choice of θ. The derivative of $J(\theta)$ can be written as:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \sum_i x_j^{(i)}(h_\theta(x^{(i)}) - y^{(i)}) \tag{4.14}$$

If this is written in its vector form, the entire gradient can be expressed as:

$$\nabla\theta J(\theta) = \sum_i x^{(i)}(h_\theta(x^{(i)}) - y^{(i)}) \tag{4.15}$$

See the lecture notes of Andrew for a complete explanation of logistic regression [167].

4.3 Features used for classification

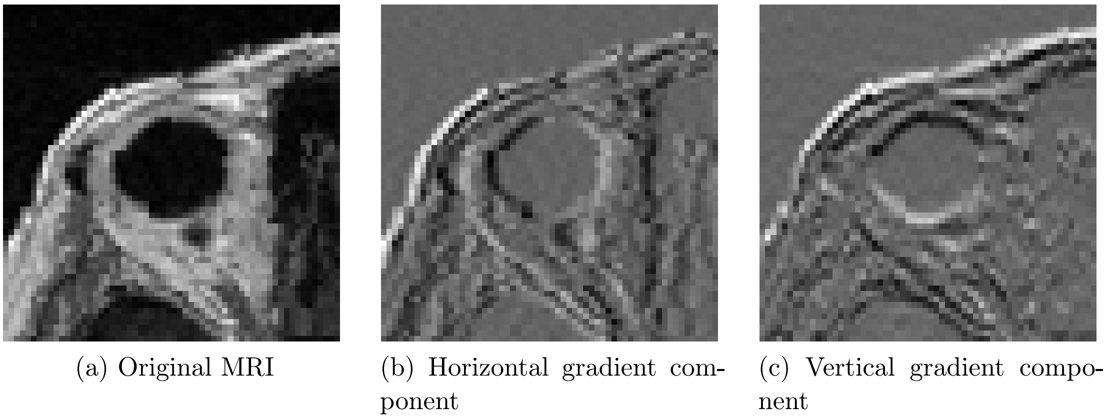
Whatever the efficacy of the machine learning strategy applied, the choice of relevant features is highly crucial on classification problems. Recent research on segmentation of brain clinical structures by machine learning techniques has tended to focus on the use of several learning algorithms rather than in the addition of more discriminative features into the classification scheme. Traditional features explained in Chapter 3, section 3.7.1 have been commonly employed when segmenting brain structures with a considerable success. However, the use of alternative features may (i) improve classification performance, while (ii) reducing, in some cases, the number of features used to describe the texture information of a given region. Apart from the application of SDAEs to the OARs segmentation problem, one of the main contributions of this work is the use of features that have not been previously employed to segment brain structures.

Among the full set of OARs involved in the RTP, there are some that present a sort of homogeneity in texture and variation in shape is less strong than in the other OARs. In this group we can include the brainstem, eyes and lens. Contrary, there are some other OARs which texture is more heterogeneous, shape variations across patients are more pronounced and/or its small size and localization variation makes automatic segmentation more complex. This second group is comprised by the optic nerves, optic chiasm, pituitary gland and pituitary stalk. Because of dissimilarities between characteristics of both groups, some of the suggested features are organ dependent, not being suitable for all the organs investigated in this work. While segmentation of some organs will exploit the use of the Geodesic Distance Transform and 3D-Local binary pattern to achieve better results, for example, the segmentation of some other will make use of texture and contextual analysis to improve the results.

4.3.1 Gradient and contextual features

In the image domain, the image gradient can be seen as a directional change in the intensity or color in an image. The image gradient is composed by two components: horizontal and vertical component (Figure 4.9). The horizontal component shows the variation of gray levels in an image along the horizontal direction, usually from left to right. This change is encoded in the grey level of the image showing the horizontal component. Thus, mean levels represent no change, bright levels represent variation from a dark value to a brighter value, and the dark level represents a change from a bright value to a darker one. Analogous observations can be made for the vertical component, which shows image variations in the vertical direction, in a top-to-bottom fashion. Combining both components, the magnitude and the orientation (Fig. 4.10) of the gradient can be obtained.

Although image gradient brings a more exhaustive description of an in-


(a) Original MRI

(b) Horizontal gradient component

(c) Vertical gradient component

Figure 4.9: A MRI slice of the brain showing partially the head in the eye region (a). While the horizontal gradient component in the x direction measuring horizontal change in intensity is displayed in (b) the vertical gradient component in the y direction measuring vertical change in intensity is shown in (c).

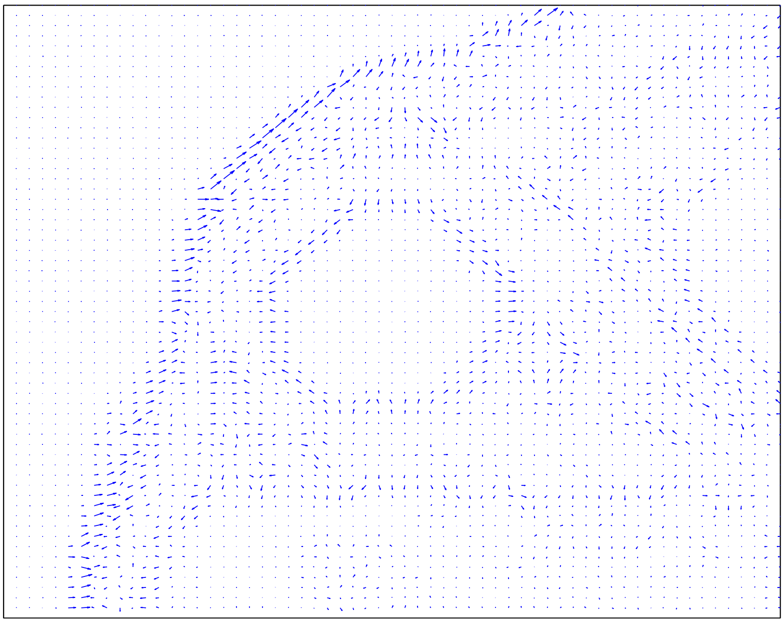


Figure 4.10: Gradient orientation values of the previous image in figure 4.9,a. The arrows indicate the direction of the gradient at each pixel.

stance, i.e. a single voxel, supplementary knowledge has been included in the features vector. This is the case of the augmented features vector. The term of augmented features vector, and the inclusion of gradient and contextual features into it, was already introduced by [168]. In their work, gradient ori-

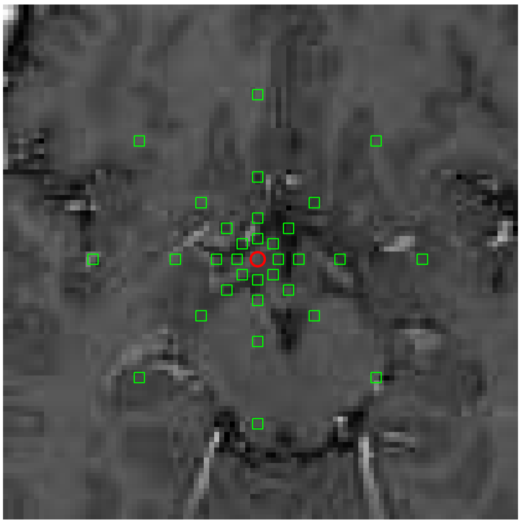


Figure 4.11: Contextual feature is defined by a number of surrounding regions (green squares) of the voxel under examination (red dot).

entations of all the voxels on each patch were used. Following their work, to describe relative relations between an image patch and its surroundings, contextual features are used. For each voxel v, a number of regions around its surroundings are sampled, radiating from voxel v with equal degree intervals and at different radius (Fig. 4.11). To obtain a continuous description of the context, intensity difference between the voxel v and a patch P is defined:

$$d_{v,P} = \mu_P - I_v \tag{4.16}$$

where μ_P is the mean intensity of the patch P and I_v is the intensity of the voxel v. In addition, a compact and binary context description is obtained by employing the Binary Robust Independent Elementary Features (BRIEF) descriptor [169]:

$$b_{v,P} = \begin{cases} 1 & I_v < \mu_P \\ 0 & otherwise \end{cases} \tag{4.17}$$

Then, for each patch, the contextual feature includes both the continuous and binary descriptor for all the neighbor regions sampled.

4.3.2 Features from texture analysis

Additionally to the information extracted from the context, texture analysis (TA) has proven to be a potentially valuable and versatile tool in neuro MR imaging [170]. MR images contain a lot of microscopic information that may not be assessed visually and texture analysis technique provides the means for obtaining this information. Therefore, we also considered the use of some these features. This is the case of statistical features of first order statistics and spectral features. TA can be divided into categories such as structural, model-based, statistical and transform, according to the means employed to evaluate the inter-relationships of the pixels. Statistical methods are the most widely used in medical images. On these methods, the spatial distribution of grey values are analyzed by computing local features at each point in the image, and deriving a set of statistics from the distributions of the local features. Local features are defined by the combination of intensities at specific position relative to each point in image. In the literature, the use of these features to characterize textures have been mainly employed for classification of images [171] or for the characterization of healthy and pathological human cerebral tissues [172]. Nevertheless, their use as discriminant factor in the segmentation of critical structures in brain cancer has not been investigated yet.

To quantitatively describe the first order statistical features of an image patch P, useful image features can be obtained from the histogram. In the proposed work the following features were employed: mean, variance, skewness, kurtosis, energy and entropy. The mean takes the average level of intensity of the image or texture being examined, whereas the variance describes the variation of intensity around the mean. Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case. Energy is a measure of local homogeneity. Energy values range from 0 to 1, where the higher the energy value, the bigger the homogeneity of the texture. Thus, for a constant image, its energy is equal to 1. Contrary, entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. It represents the opposite of the energy. A completely random distribution would have very high entropy because it represents chaos. An image with a solid tone would have an entropy value of 0.

Probability density of occurrence of the intensity levels can be obtained by dividing the value of intensity level histogram by the total number of pixels in an image:

$$P(i) = h(i)/n_x * n_y \qquad i = 0, 1, ...G - 1 \tag{4.18}$$

where n_x and n_y are the number of pixels in the horizontal and vertical image domain, respectively. G represents the total gray levels on the image. Thus, features obtained from the histogram are calculated as follows:

$$Mean: \qquad \mu = \sum_{i=0}^{G-1} ip(i) \tag{4.19}$$

$$Variance: \qquad \sigma^2 = \sum_{i=0}^{G-1} (i-\mu)^2 p(i) \tag{4.20}$$

$$Skewness: \qquad \mu_3 = \sigma^{-3} \sum_{i=0}^{G-1} (i-\mu)^3 p(i) \tag{4.21}$$

$$Kurtosis: \qquad \mu_4 = \sigma^{-4} \sum_{i=0}^{G-1} (i-\mu)^4 p(i) \tag{4.22}$$

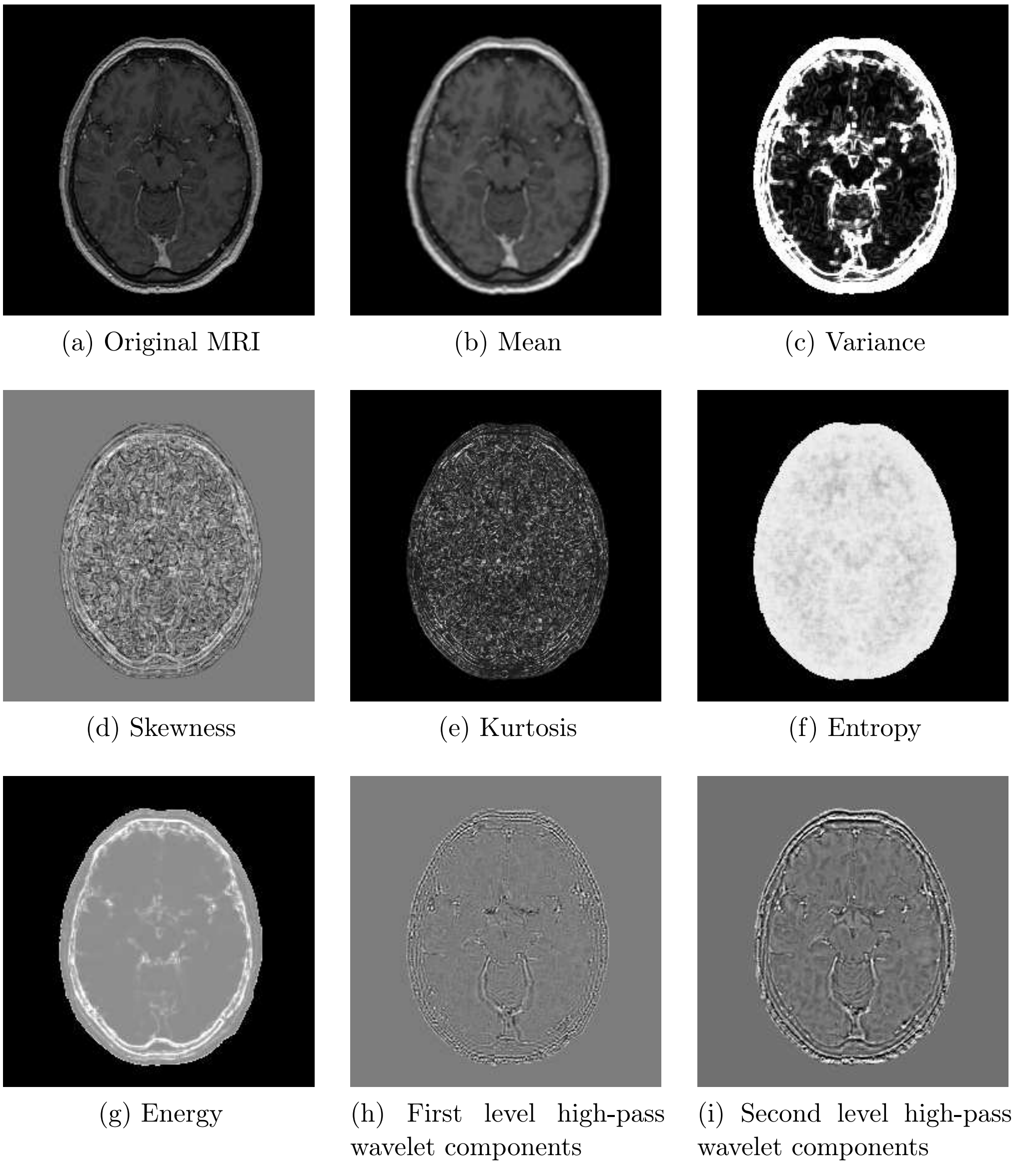
$$Energy: \qquad E = \sum_{i=0}^{G-1} [p(i)]^2 \tag{4.23}$$

$$Entropy: \qquad H = -\sum_{i=0}^{G-1} p(i) log_2[p(i)] \tag{4.24}$$

Statistical based features may lack the sensitivity to identify larger scale or more coarse changes in spatial frequency. To evaluate spatial frequencies at multiple scales wavelet functions can be employed [173]. The basic idea of the algorithm is to divide the input images into respective decomposed sub-images using the wavelet transform. A wavelet transform decomposes a signal to a hierarchy of sub-bands with sequential decrease in resolution. The idea of using the wavelets to extract information in texture classification context is not entirely new. Specifically, in the medical field, Discrete wavelet transform (DWT) has been used for sub-domains such as image fusion, image resolution enhancement or image segmentation [174]. Despite this, a major usage of DWT has been noticed for classifying MR brain images into normal and abnormal tissue [175].

4.3.3 Geodesic Distance Transform Map

To encourage spatial regularization and contrast-sensitivity, geodesic distance transform map (GDTM) of the input image is used as additional feature. The addition of GDTM in the features vector used by the classifier exploits the ability of seed-expansion to fill contiguous, coherent regions without regard to boundary length. As explained in the work of Criminisi et al. [176], given an



(a) Original MRI

(b) Mean

(c) Variance

(d) Skewness

(e) Kurtosis

(f) Entropy

(g) Energy

(h) First level high-pass wavelet components

(i) Second level high-pass wavelet components

Figure 4.12: First-order statistical features (F-OSF) example. Axial slice of a brain with several first-order statistical features computed with a with radius = 3 around each voxel. First and second levels of high-pass components from wavelets decomposition ((h) and (i)).

image I defined on a 2D domain ψ, a binary mask M (with $M(\mathrm{x}) \in \{0,1\}\ \forall \mathrm{x}$) and an "object" region Ω with $\mathrm{x} \in \Omega \iff M(\mathrm{x}) = 0$, the unsigned geodesic distance of each pixel x from Ω is defined as:

$$D(x; M, \nabla I) = \min_{\{x' | M(x')=0\}} d(x, x'), \qquad with \tag{4.25}$$

$$d(a,b) = \min_{\Gamma \in \mathcal{P}_{a,b}} \int_0^1 \sqrt{\|\Gamma'(s)\|^2 + \gamma^2(\nabla I \cdot u)^2 \ ds} \tag{4.26}$$

with $\mathcal{P}_{a,b}$ the set of all paths between the points **a** and **b**, and Γ(s) : $\Re \to \Re^2$ indicating one such path, which is parameterized by $s \in$ [0,1]. Figure 1 shows an example of how compute the GTDM of an image given a binary mask.

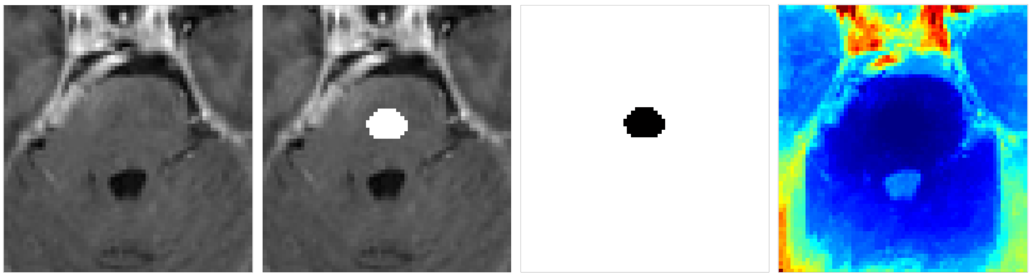


Figure 4.13: Geodesic distance transform map: a) axial MR view of the brainstem, b) mask obtained from the probability brainstem map (in white), c) binary mask used to obtain the GDTM, and d) output GDTM.

4.3.4 3D Local Binary Texture Pattern

In order to catch neighborhood appearance of the voxel under examination with the fewest number of features, Local Binary Patterns (LBP) are investigated. The idea of LBP is to give a pattern code to each voxel. Particularly, an extended version of 3D-LBP presented by [177] (Fig. 4.14) is proposed. In their work, classical LBP [178] were adapted by selecting the 6 nearest voxels and ordering them to create the encoding patterns (Figure 2). By encoding patterns in that manner, $2^6 = 64$ possible patterns would be created. However, those 64 possible combinations were merged in 10 different groups according to geometrical similarities (Figure 4.14). In accordance with this classification, each group is filled with patterns that have the same number of neighbor voxels with a gray level higher than the central voxel c. Thus, rotation invariance in each group is kept. These groups are defined with (Table 4.2):

$$card(c) = \sum_{i=0}^{P-1} s(g_i - g_c) \tag{4.27}$$

where P = 6 is the number of neighboring voxels and R=1 or R=2 the distance between central voxel c and its neighbors i. By using R =1,2 micro and macro-structure appearance of the texture are captured in the 3D-LBTP. In equation 3, card(c) gives the number of neighbors with a higher gray level than the central voxel c.

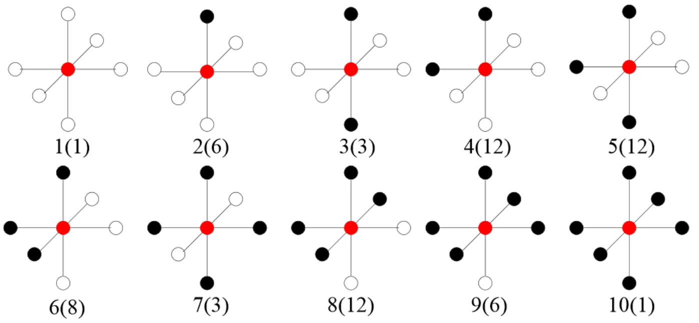


Figure 4.14: Merging the 64 possible patterns into 10 groups. Number of different patterns for each group is indicated in brackets.

LBP3D	card(c)	Condition
1	0	
2	1	
3	2	opposite voxels
4	2	bend voxels
5	3	voxels on the same plane
6	3	voxels on different planes
7	4	voxels on the same plane
8	4	voxels on different planes
9	5	
10	6	

Table 4.2: Definition of the 10 groups of patterns.

In addition to the encoded value for the 3D patch structure proposed by [177], an additional texture value is included. Let g_{high} the gray values that are higher than the gray value of the center voxel c in the 3D-LPB (Figure 2). Similarly, let's denote g_{low} to the gray values that are lower than the gray value of the center voxel c in the 3D-LPB. Then, the texture value added to the encoded structure value is defined as:

$$Texture_{val} = mean \sum_{i=0}^{m} g_{high}(i) - mean \sum_{i=0}^{n} g_{low}(i) \tag{4.28}$$

where m and n are the number of neighboring voxels with higher and lower values than the center voxel c, respectively. Thus, the introduction of the 3D-LBTP in the features vector will lead to 4 new features: 3D-LBP and Texture$_{val}$ for R = 1 and 2.

4.4 Training the deep network

This section presents the way we combine the deep network with the proposed features. First, pre-processing required for the images to be used in this work is explained. Next, training and classification of the network are detailed.

4.4.1 Pre-processing

Pre-processing involves any of the diverse processes that help the segmentation algorithm to produce a more accurate model. Ideally, the segmentation process should be fully automatic, not requiring any user interaction. Nevertheless, this barely happens. Typical pre-processing steps include registration of images to a common coordinate space, intensity normalization, resampling of images to the same resolution or the bias field correction, for example. Only pre-processing methods applied to the images in this thesis are explained above.

4.4.1.1 Resampling

MR resolution is not always the same. Particularly, differences in resolution often come from the x and y coordinates. Hence, to make both the training and classification more homogeneous, images which resolution differed from 1mm x 1mm x 1mm were resampled to this resolution.

4.4.1.2 Patient Alignment

If the whole set of images in a study are first aligned to a common template, a specific region of interest is then already in approximately the same region of the coordinate space for all subjects across the study. This fact makes learning patterns easier and reduces the search space for a particular region of interest. It is therefore a common practice in brain segmentation approaches to apply some sort of registration technique to the MRI images to make them as similar as possible to a common MRI template. Some approaches require a rigid registration step to align the images [128,138]. However, in the proposed approach, and as in [73] and [76], MRI T1 images were spatially aligned such that the anterior commissure and posterior commissure (AC−PC) line was horizontally oriented in the sagittal plane, and the inter hemispheric fissure was aligned on the two other axes. This process therefore represents the initialization step for the segmentation of a new target patient.

It is worthwhile to describe the coordinate system used to define neuroanatomical locations of normalized images. The 'Talairach' coordinate system specifies locations relative to their distance from the anterior commissure

(AC). The AC is a thin white matter tract between the olfactory areas of each hemisphere, which despite of being a small region represents an easy spot to localize, making of it an ideal origin for the coordinate system. Each location is described by three numbers, each describing the distance in millimeters from the AC: X is the left/right dimension, Y is the posterior/anterior dimension, and Z is the ventral/dorsal dimension. The diagram below shows the location of the AC (blue dot) on a midsagittal view. Note that the orientation of axial plane in Talairach space officially lies immediately dorsal to the AC and ventral the posterior commissure (PC, yellow dot), as in Figure 4.15.

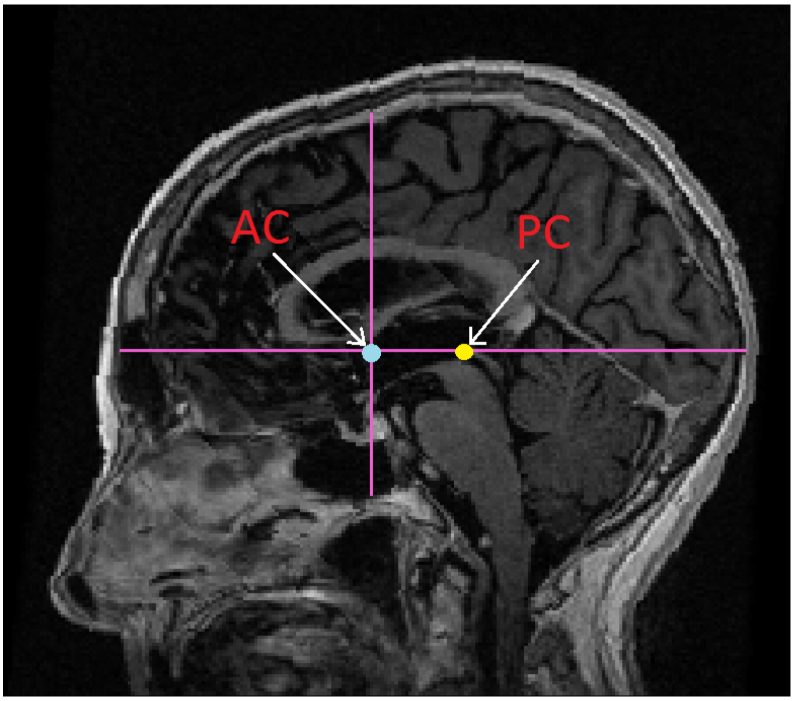


Figure 4.15: AC-PC example on MRI image(right).

4.4.1.3 Image Normalization

Image normalization is a process that changes the range of pixel intensity values. Normalize an image by setting its mean to zero and variance to one. To do so, images intensity values are shifted and scaled so that the voxels in the image have a zero mean and unit variance. The filter *NormalizeImageFilter* from the Insight Segmentation and Registration Toolkit (ITK) [179] was used to normalize the images.

4.4.2 Training

Supervised learning based approaches involve the existence of two distinct groups of images: training and testing images. The first group is composed by images that have been manually segmented by experts. Manually labeled images are often referred to as *reference* or *standard contours*. This set of images, comprising both clinical images and manual labels, are utilized to learn patterns associated with a particular structure. On the other hand, the testing images group is composed by an independent set of images, which are not included in the training set. Testing images are used to validate how well the patterns were learned. To compare an algorithm's performance with the *reference contour* defined by an expert, the testing images must also be manually segmented.

The whole process used in the training step is shown in figure 4.16. The first step in the training consists on creating a common binary mask for each of the OARs. This mask was computed by applying an "or" operation to all the reference masks in the training set for a given OAR. This mask was employed to prune the voxels in all the images, both in training and classification, and thus reducing the research region of each OAR. Therefore, only voxels allocated inside the common mask are taken into account when extracting the features that will be used in the classifier. Once all the features are extracted, scaling is applied over all of them. At last, the training model is computed for the desired classifier.

4.4.2.1 Probability map and common mask creation.

A detailed example of how the probability map and the common mask are created during the training phase is shown in Fig.4.17. Masks contained in the training set are added into a volume to create a probability map for each OAR, which yielded voxel-wise continuous probabilistic measures in the range of [0,1], indicating the likelihood of the organ class. This map represents the frequency with which an OAR appears in the training set and therefore the probability of a given voxel to belong to some structure. The probability map is also used to reduce the number of samples that are fed into the classifier. From this map, a region of interest (ROI) mask is generated. The pruning criterion is based on the probability of a voxel to belong to any of the structures of interest. Thus, any voxel containing a probability higher than zero is taken into account to create the common mask, for each structure, which will be used to prune the voxels in the feature extraction stage. To ensure that OARs of unseen patients will be inside this common mask a security margin was given to the generated mask by applying a morphological dilation.

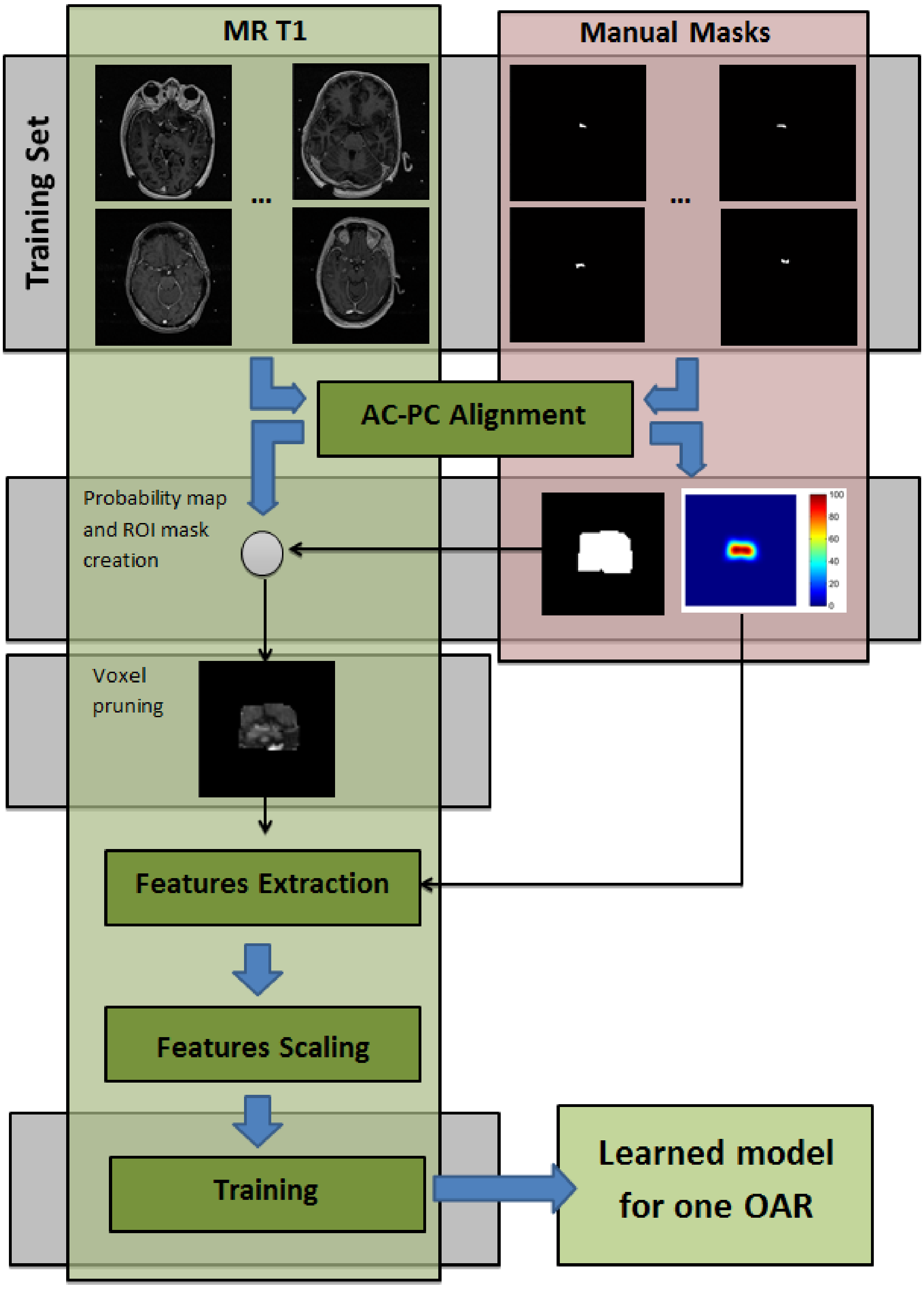


Figure 4.16: Framework for training.

4.4.2.2 Features extraction

Traditional features used in the segmentation of brain structures were already introduced in Section 3.7.1. In addition to these ones, new proposed features

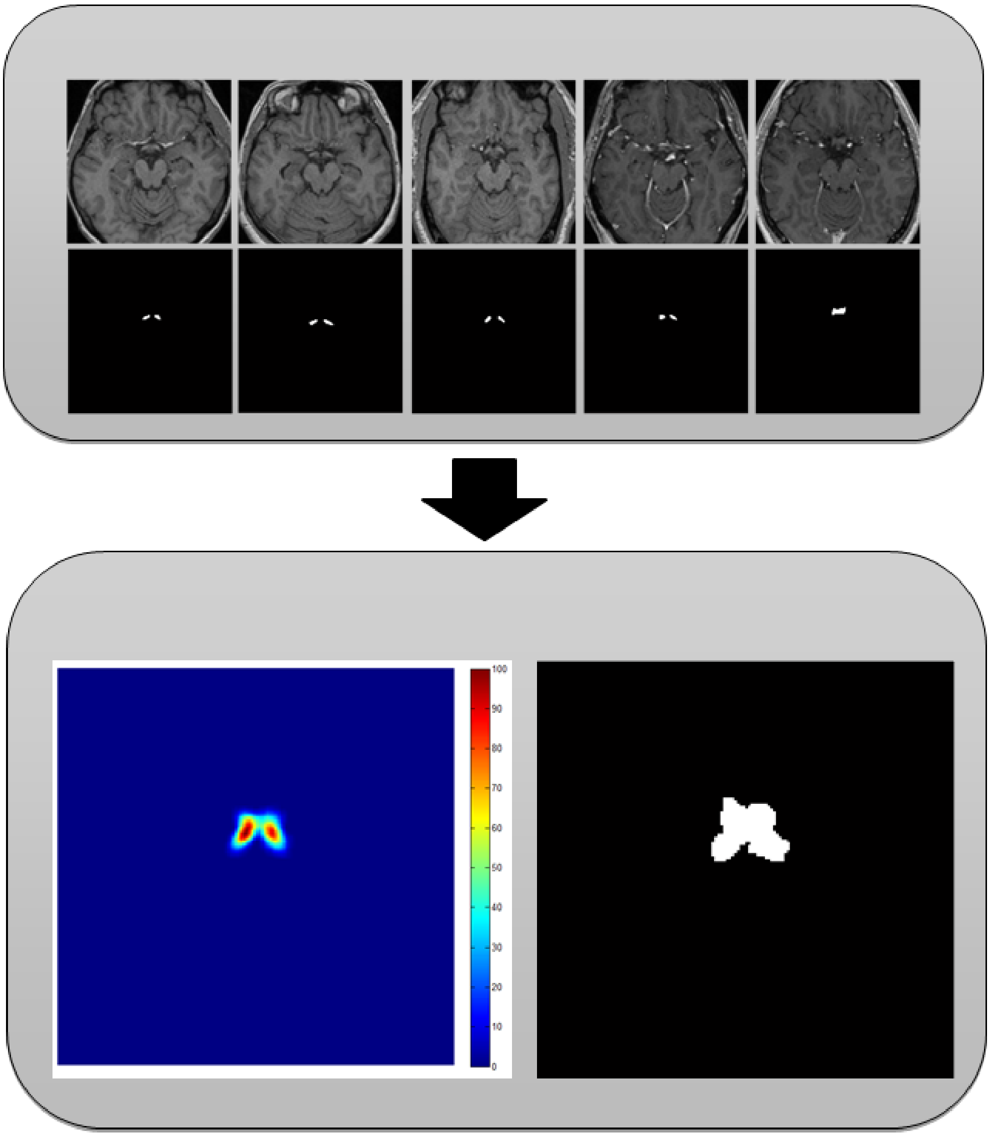


Figure 4.17: Probability map (*bottom-left*) and common mask (*bottom-right*) creation process. Example showing a 2D axial slice of the optic chiasm.

were detailed in Section 4.3. However, it is important to note that features employed in the classification may slightly vary from one organ to another, depending on some characteristics of the organs to segment. Thus, for example, the use of additional spatial information, such as distance to the center of the brain, and angle with respect to the horizontal can help to the segmentation of symmetric thin structures, such as the optic nerves. In the other hand, the use of features that encourage spatial regularization over the entire structure improves the classification in large and/or well-defined structures, such as the brainstem or the eyes. We can say, therefore, that each structure requires its own specific descriptors.

Features are extracted on voxels that belong to the inner part of the ROI mask defined in previous section. Hence, we avoid to analyze voxels which do not give any relevant information to solve our problem.

4.4.2.3 Scaling

As explained in [180], scaling the features before applying non-scale invariant techniques, such as SDAE, is very important for a good performance of the classifier. Among the main advantages of scaling, it can be mentioned that it helps to: 1) avoid attributes in greater numeric ranges dominating those in smaller numeric ranges, and 2) avoid numerical difficulties during the calculation. Complications in the calculations can be caused by large attribute values when the inner products of feature vectors are used to compute the kernel values.

Ranges $[-1, +1]$ and $[0, +1]$ are typically employed to scale the attributes of the features vectors. Range selected to scale training data must be coherent with range used to scale the testing data. This means that the same scaling factors must be used for both training and classification data and not scale them separately. Let's imagine that we have a features vector of intensities with 8 attributes indicating grey levels in the training that we scale in the range $[0, +1]$ (second row of table 4.3). For a given features vector on the testing, if scaling is done independently of the data contained in the training set (fourth row of table 4.3), the scaled values are not correlated with those in the training. As a consequence, the classification performance will be unsatisfactory in comparison with features correctly scaled (row five of table 4.3). In appendix B of [181] a real example showing differences in classification accuracy between wrong and right scaled values is detailed.

	Features vector elements							
	#1	#2	#3	#4	#5	#6	#7	#8
Training (Original)	178	205	189	35	12	48	255	241
Training (Scaled)	0.6980	0.8039	0.7412	0.1373	0.0471	0.1882	1.0000	0.9451
Testing (Original)	201	198	55	33	45	124	89	174
Testing (Erroneously Scaled)	1.0000	0.9821	0.1310	0	0.0714	0.5417	0.3333	0.8393
Testing (Correctly Scaled)	0.7882	0.7765	0.2157	0.1294	0.1765	0.4863	0.3490	0.6824

Table 4.3: Scale example showing a bad and a good example of features scaling

4.4.2.4 Parameters setting of the classifier

Performance of classification algorithms also depends on the selection of their parameters, which must be carefully selected by the user. However, suitable combination of parameters depends on the training data. As a result, the parameters choice of the different classifiers can be viewed as an optimization process, where parameters values are iteratively modified until a satisfactory result is achieved. Best parameters may be selected by users based on a priori knowledge and\or expertise [148]. Nevertheless, this often implies the user to manually test a wide range of parameters and select the best combination of them, which is time-consuming. Additionally, a risk of overfitting still prevails when different settings for the classifiers are evaluated. Parameters can be adjusted until the classification optimally performs, allowing a ”leakage” of knowledge about the testing set into the model which would no longer provide a generalization on its performance. To tackle these issues some validation techniques for model selection have been adopted. Next section introduces the use of cross-validation to select a successful combination of the classifier's parameters.

4.4.2.4.1 Cross-validation for model selection

In this section we consider how to use methods of cross-validation (CV) for model selection. The parameters of a classifier have to be optimized based on the training available data. An independent testing set is therefore required for making a reliable assessment of the applicability of the classifier to new data. Cross-validation provides a simple way to measure this generalization performance when no such test data are available. A common strategy is to separate the training data set into two disjoint sets. One of these sets is actually used for training, and the other, the validation set, which is used to monitor the performance. The prediction accuracy obtained from the unknown set more precisely reflects the performance on classifying an independent data set. The performance on the validation set is used as a proxy for the generalization error and model selection achieved using this measure.

In practice, a shortcoming of hold-out method is that only a fraction of the full data set can be used for training. In addition, if the validation set is small, the performance obtained might have large variance. To minimize these problems, CV is very often used in the k-fold cross-validation setting: the k-fold cross-validation data is split into k disjoint, equally sized subsets. Validation is then done on a single subset and training is done using the union of the remaining (k-1) subsets. The entire procedure is repeated k times, each time with a different subset for validation. Thus, a large fraction of the data can be used for training, and all cases appear as validation cases. The price is that k models must be trained instead of one. Typical values for k are in the range 3 to 10, whereas 10-fold CV has been shown to be accurate enough for model selection [182].

The following subsection highlights and details the task of parameters selection for the SDAE approach followed in this thesis. Parameters selection for SVM are detailed in Appendix B.

4.4.2.4.2 SDAE Parameter Setting

One of the most crucial, and at the same time most complex decisions to make when working with any kind of neural networks is the architecture configuration. This comprises the choice of the depth of the network, as well as the number of hidden units in each layer. Trying to find the best network configuration by performing a grid search becomes much harder than in the case of the SVM, where only two parameters were searched. The strategy followed to find a suitable network structure was based on the error convergence during training. Thus, the faster the convergence and the lower the error, the more suitable the network structure. In order to constrain the search and avoid having to test hundreds or even thousands of different network architectures, typical network configurations were employed, where the size of layer *l+1* is half of the precedent layer *l*.

In SDAE there are another parameters that must be carefully selected. These parameters include layer-wise learning rate, the activation function and the corruption level of the denoised autoencoder.

4.5 Classification

Classification is done at one class at each time. That means that a binary classifier is used for each of the structures. In this context, classes for each classifier are: one structure of interest and the background. Classification scheme, although very similar, is slightly different from the scheme used during the training phase. In figure 4.19 the pipeline followed to segment a new patient, or target patient, is presented.

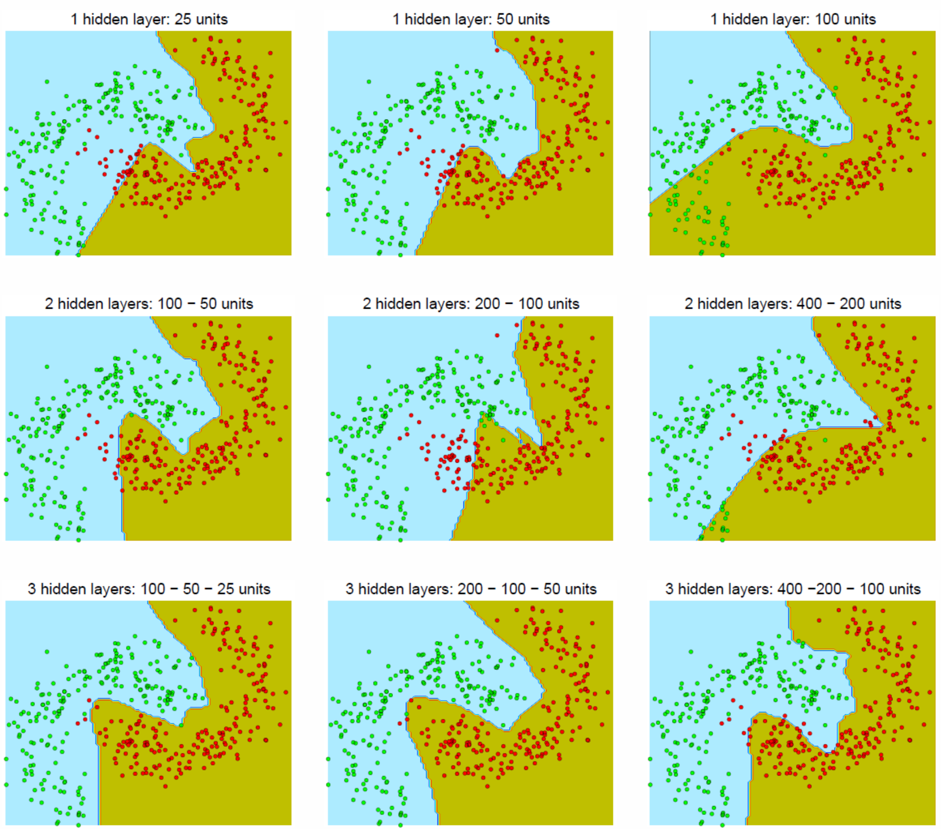


Figure 4.18: Decision boundaries for a banana shaped dataset generated by SDAE with different network architectures.

4.5.1 Pre-processing

Pre-processing steps required for classifying a target patient are the same than those presented on the training section (Section 4.4.1). These steps are: resampling, patient alignment and image normalization.

4.5.2 Features extraction

Voxel pruning is done with the common mask generated during the training (section 4.4.2.1) for each new target patient and each OAR. Then, features to be used in the classifier are extracted from voxels inside each ROI. As happened in the training phase, features will slightly vary from one organ to each other. However, for the same organ, features composing the features array are the same both in training and classification.

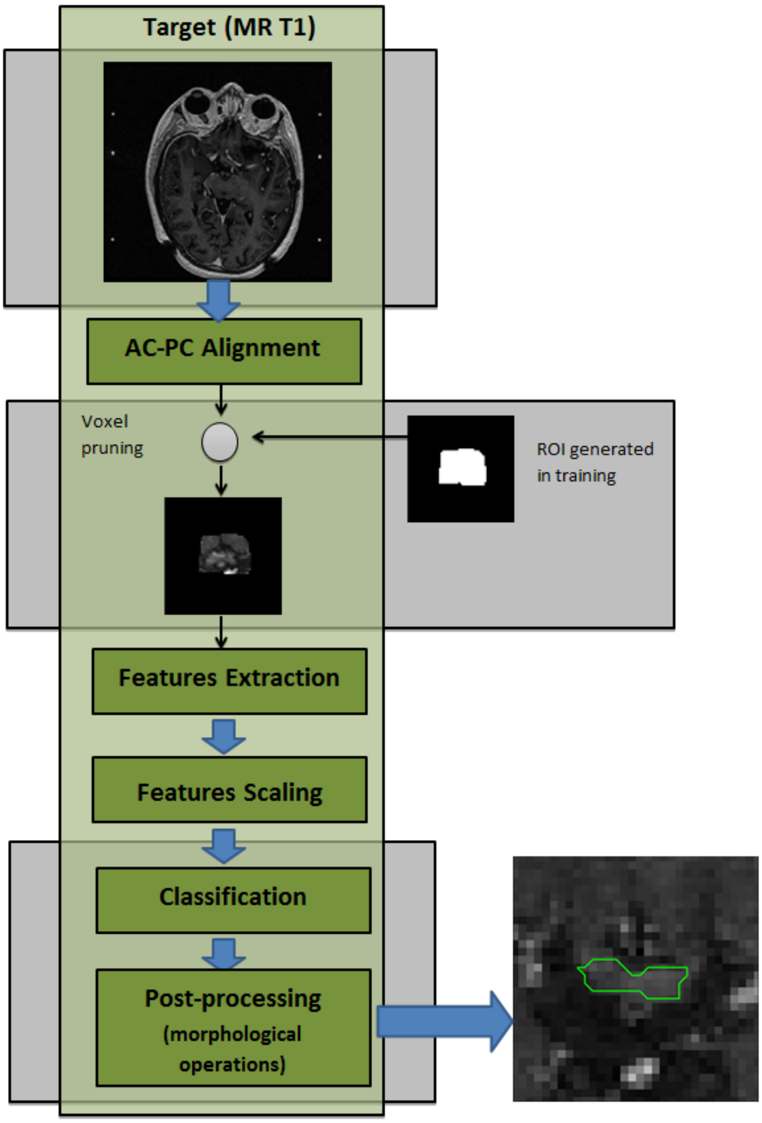


Figure 4.19: Framework for classification.

4.5.3 Scaling

As stated in Section 4.5.3, features extracted for classification are scaled in concordance with scaling values used during the training phase. Using different scaling values will negatively affect the segmentation performance.

4.5.4 Classification

Classification basically consists on applying the weights learned during the training stage to each input sample. Thus, once features for all samples have been extracted and scaled, they are fed into our trained network. Input features are multiplied by learned weights from the first layer. Output from first layer is multiplied by weights on the second layer. This process is repeated until the last layer, which gives a value indicating whether the sample belong to the OARs class.

4.5.5 Post-processing

After classification, a post-processing layer, which was mainly a filter applying morphological operations was introduced before providing the output. Particularly, a closing operation to remove small isolated regions and to fill small holes was employed.

Chapter 5
Materials and Methods

"The best time to plant a tree was 20 years ago. The second best time is now."
Chinese Proverb

In this chapter the materials employed to conduct this work, as well as to evaluate the performance of the proposed approach are presented. First section introduces the software used to develop all the content of this thesis. Then, imaging data employed on the experiment is presented. Medical imaging analysis, and particularly segmentation, often lacks from a universal ground truth. Thus, multiple observers are typically required to manually delineate a set of structures on a group of patients, from which reference contours can be therefore generated. This second section details the process followed to generate the reference standard. Third section details the evaluation metrics employed to analyze results and how important they are for the assessment of our proposed classification scheme in clinical context. To evaluate whether there exist significant differences between groups, statistical analysis are often employed. This type of analysis is described in last section.

5.1 Software

All the code that has been employed in this thesis has been implemented using the following platforms: MATLAB(The MathWorks Inc., Natick, MA, 2000) and Microsoft Visual Studio (MSVS) 2010.

There are two main processes in the code developed in this thesis: image processing step (i.e. features extraction) and learning/classification. For the former step, a whole set of functions were developed in MSVS 2010 by using C++ programming language. The learning and classification steps for the deep networks were implemented on MATLAB based on the toolbox provided by Palm [156]. The publicly available library libsvm [183] was used to compare our classification scheme with SVM.

Apart from the research contribution provided by this work, we aim at developing some prototype, that is why we also employed MSVS. The main program run on this platform. To connect MSVS with the MATLAB functionalities of the deep learning toolbox, the MATLAB run-time compiler was employed to create the dynamic libraries (i.e. dlls) to be included in the MSVS project. Thus the whole process is as follows:

1. Features extraction is performed by employing functions implemented in C++.
    - All features have been extracted.
    - An array containing all the features is created.
2. The MATLAB dll that contains the deep learning functionalities is called from MSVS.
    - Either training or classification is performed.
    - According to the operational mode (training/classification) some information is received (trained model or an array containing the predicted labels).
3. The segmentation is reconstructed by employing C++ code.

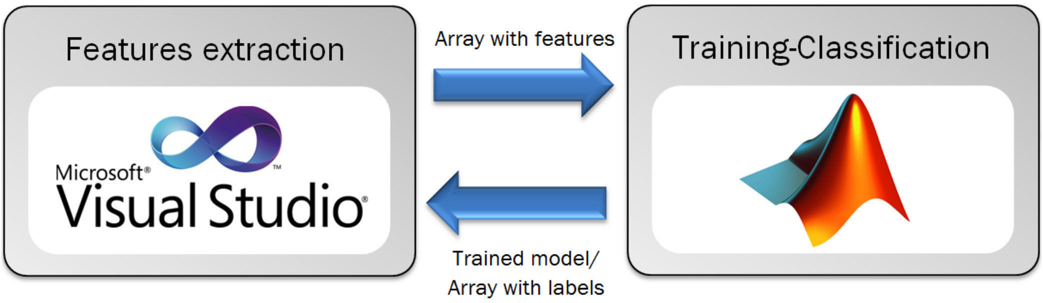


Figure 5.1: Workflow of the connection between MSVS and MATLAB.

For the manual labeling, Artiview 3.0 (AQUILAB) was used by the observers that participated in the study of this thesis.

5.2 Method validation

Validation of medical image processing methods is of crucial importance because the performance of such methods can have an impact on the performance of the larger systems in which they are embedded. Definition of a standard protocol for validation may therefore have a high relevance to facilitate the complete and accurate reporting of validation studies and results and the comparison of such studies and results. Following the guidelines suggested by Jannin et al. [184] towards this standardization we designed the validation protocol of our method.

5.2.1 Validation objective

In the clinical context of segmentation of organs at risk of brain cancer patients undergoing radiotherapy or radio surgery, a segmentation method based on a stack of denoised auto-encoders fed by a wide range of image-based features extracted from MR-T1 images is able to segment those organs at risk with an accuracy that is significantly better than other state-of-the-art methods and that lies between experts variability.

5.2.2 Validation process

The validation process is performed on a validation dataset, which detailed description is of high importance. Image data employed in this experiment was composed by clinical images, which description is presented in Section 5.3.1. Given the validation datasets, the outcome of the segmentation method has to be validated. The segmentation method computes an estimate of the reference standard, being the reference standard the theoretical ideal result. In this work, the reference was provided by expert observers (Sections 5.3.2 and 5.3.3). By comparing outcomes of the segmentation method and reference standard, a validation criterion aims at characterizing different properties of the method to be validated. These properties may include accuracy, robustness or efficiency, for example. Evaluation metrics employed in this work to validate our segmentation method are introduced in Section 5.5. To do these comparisons, output volumes from the segmentations are used. It is commonly to compare results from a proposed method against a well known state-of-the-art method. In our case, support vector machines (SVM) was the approach chosen for comparison purposes. The last part of the validation process comprises the analysis of results (Section 6.2). First, results computed by the proposed segmentation method and the reference method, i.e. SVM, are compared. Segmentation results are also compared against manual annotations. Then, comparison results are tested against the validation hypothesis (Section 5.2.2) in order to provide the validation result.

5.3 Imaging Data

5.3.1 Dataset

MRI data from 15 patients who underwent Leksell Gamma Knife Radiosurgery were used in this work. Two different MRI facilities were employed to acquire images according to the radiosurgery planning protocol (Table 5.1). Pathologies in this dataset included trigeminal neuralgia, metastases, and brainstem cavernoma. Although the employed dataset was limited in size, it was representative of the population. Examples of the original input sequences from several patients are shown in Figure 5.2. In this figure, axial slices showing some tumors on these patients are presented.

Experiments were retrospectively performed on all the patients. All data analyzed was collected as part of routine diagnosis and treatment. Prior to being processed all images were anonymized. Patients were diagnosed and treated according to national guidelines and agreements. Therefore, no consent approval for our study was required.

MRI System	TE(ms)	TR(ms)	Echo number	Matrix size	Seq. Name	Voxel Size (mm^3)
Philips Achieva 1.5T	4.602	25	1	256x256	T1 3D FFE	1x1x1
GEHC Optima MR450w 1.5T	2.412	5.9	1	256x256	FSPGR	0.8203x0.8203x1

Table 5.1: Acquisition parameters on the 2 MRI devices.

Figure 5.3 shows the intensity profile of some OARs for a given patient. From this image, it can be seen that structures share intensity bands between them, which makes no possible to only employ voxel intensity values to separate them. In addition, some properties of the OARs across the patients included in this study are presented in Table 5.2.

Image characteristics

	Intenstiy			**Volume**
	Mean	**Max**	**Min**	**Size (cm^3)**
Brainstem	280.62 (± 277.23)	745.67 (± 632.45)	36.56 (± 34.50)	25.79 (± 2.85)
Eye (Right)	117.74 (± 67.04)	542.39 (± 305.95)	3.31 (± 2.78)	5.41 (± 0.73)
Eye (Left)	118.68 (± 78.49)	539.92 (± 292.17)	3.69 (± 3.24)	5.43 (± 0.78)
Lens (Right)	438.38 (± 272.19)	619.93 (± 381.16)	233.71 (± 140.01)	0.15 (± 0.04)
Lens (Left)	438.56 (± 289.95)	640.31 (± 434.34)	257.43 (± 159.29)	0.14 (± 0.06)
Optic nerve (Right)	480.81 (± 286.01)	959.14 (± 539.85)	86.14 (± 91.01)	0.81 (± 0.18)
Optic nerve (Left)	484.89 (± 294.52)	994.79 (± 626.46)	94.57 (± 125.21)	0.82 (± 0.25)
Optic chiasm	497.25 (± 324.11)	734.53 (± 523.55)	262.50 (± 168.85)	0.23 (± 0.05)
Pituitary Gland	748.85 (± 478.79)	1210.15 (± 736.49)	313.21 (± 271.75)	0.53 (± 0.14)
Pituitary Stalk	568.45 (± 414.06)	939.71 (± 665.71)	295.93 (± 237.83)	0.08 (± 0.02)

Table 5.2: Intensity and volume characteristics of images contained in the dataset used for this work.

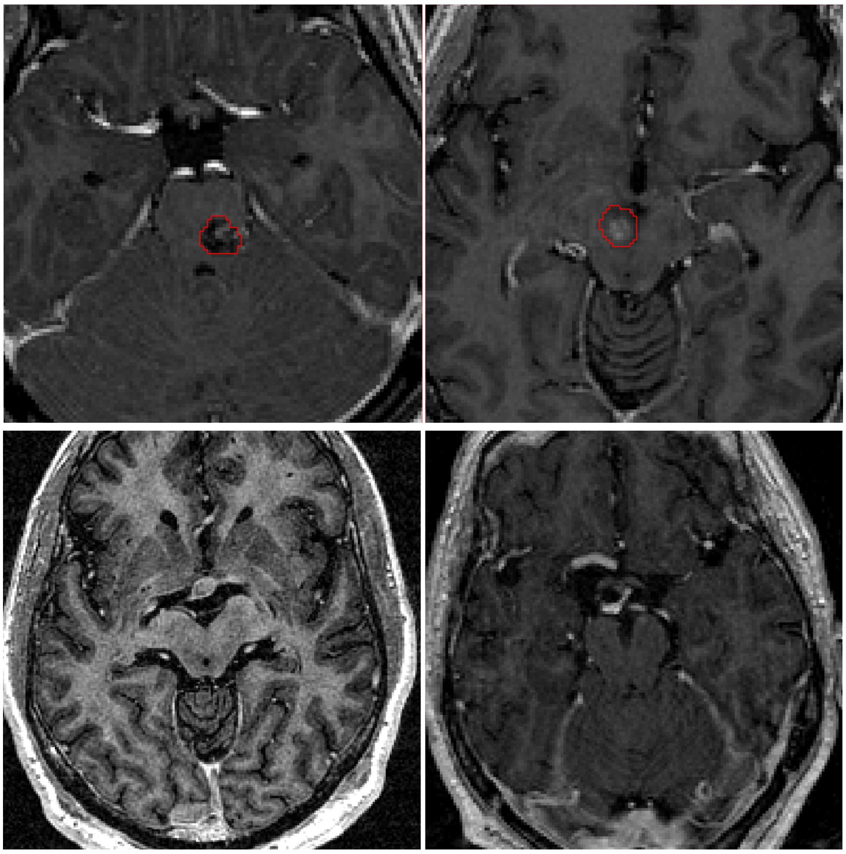


Figure 5.2: Some examples of images contained in the database employed in this work. While in some cases the tumor is inside the brainstem and may change the shape and intensity properties of the brainstem (top-row), in other brain cancer cases, tumors do not affect the brainstem or other OARs properties.

5.3.2 Manual Contouring

Altogether, four experts participated in this experiment. This group of experts was comprised by: two neurosurgeons, one physician and one medical physicist. All of them were trained and qualified for radiosurgery delineation. However, the number of available manual contours differed from one OAR to each other. Thus, the composition of the manually labeled dataset, per patient, was: four manual contours of the brainstem in 9 patients, three manual contours of the optic nerves, optic chiasm, pituitary gland and pituitary stalk in 15 patients, and only one manual contour of the eyes and lenses in 15 patients. The reason for having only one manual contour per patient for the eyes and lenses is because they do not represent a complex structure to segment. Thus, less inter-observer variation is expected, being meaningless to employ several contours to generate a reference standard. Protocol for delineation

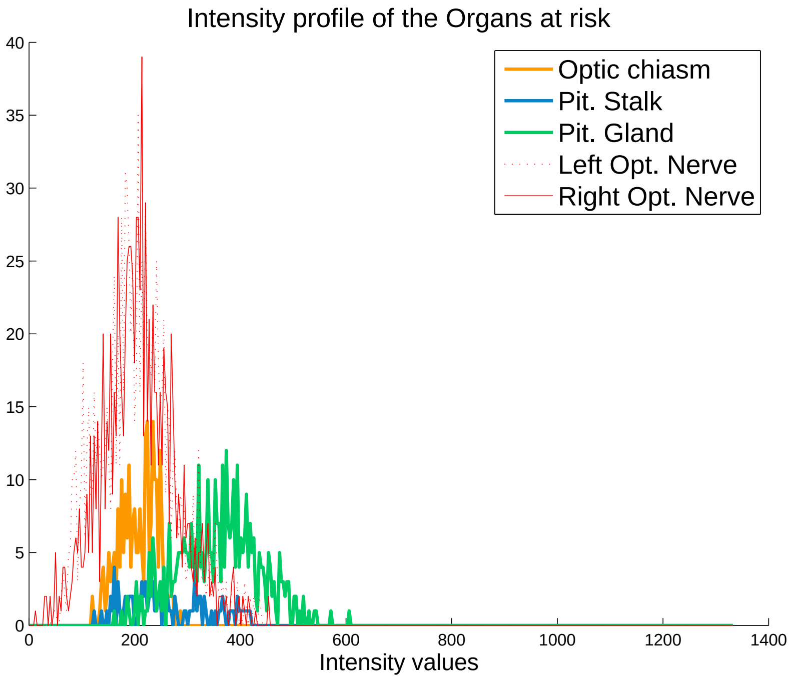


Figure 5.3: Intensity profiles of some OARs for a randomly selected patient.

was described before contouring session. Artiview ®3.0 (Aquilab) was used after a training session to achieve Dicom RT contouring structures. Average manual segmentation times per organ are listed in table 5.3.

	Manual segmentation time (minutes)
Brainstem	$20' \, 12''$ ($\pm$ $10' \, 48''$)
Eyes	$6' \, 51''$ ($\pm$ $1' \, 42''$)
Lenses	$2' \, 17''$ ($\pm$ $0' \, 51''$)
Optic nerves	$7' \, 34''$ ($\pm$ $2' \, 53''$)
Optic chiasm	$1' \, 52''$ ($\pm$ $0' \, 38''$)
Pituitary Gland	$3' \, 8''$ ($\pm$ $0' \, 55''$)
Pituitary Stalk	$2' \, 41''$ ($\pm$ $0' \, 49''$)

Table 5.3: Mean manual segmentation times per observer and organ.

The pie chart in 5.4 represents the total time for manual segmentation averaged over all the patients. The sections show the time for the OARs delineated. Looking at the section, it can be observed that brainstem, eyes and optic nerves represented the structures where the experts spent more time in the segmentation task.

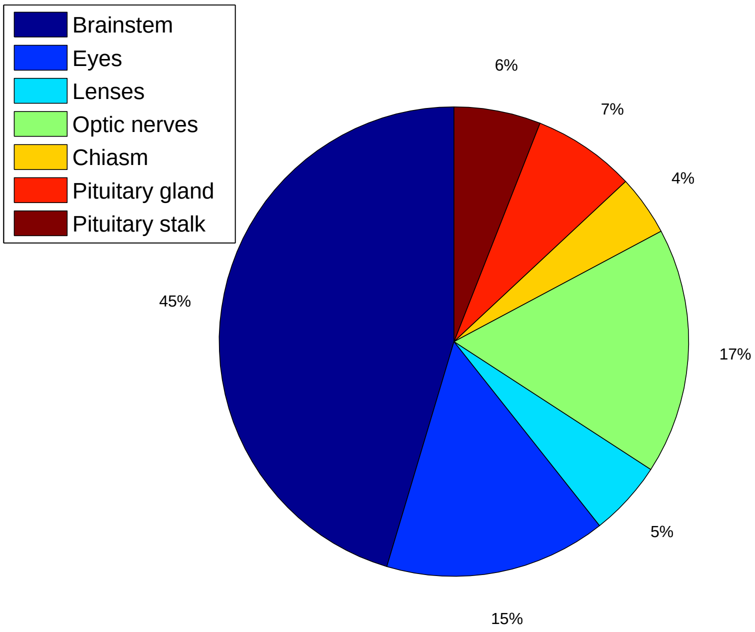


Figure 5.4: Pie charts representing mean manual segmentation times for OARs.

5.3.3 Simulated Ground Truth

To conduct a validation analysis of the quality of image segmentation, it is typically necessary to know a voxel-wise reference standard. Nevertheless, image segmentation in the medical domain often lacks from a universal known ground truth. Even though a single manual rater provides realistic data, contours may suffer from intra- and inter-observer variability. Thus, a number of observers and target patients that provide a good statistical analysis is often required. Accordingly, this study has been designed to quantify variation among clinicians in delineating OARs and to assess our proposed classification scheme in this context.

Therefore, available manual contours from the experts were used to create the simulated ground truth, which will be onwards referred to as reference. Reference contours have been obtained in this thesis by using the computationally simple concept of probability maps. In this method, which is analogous to the voting rule approach, probability maps are thresholded at a variable level in order to create the mask. The threshold was fixed at 50%, or at 75%, depending on whether the number of available manual contours from the physicians was three or four, respectively. Hence, reference contours for big structures such as the brainstem will be generated by thresholding the probability map at 75% of the maximum level. For small structures, however, threshold level will be fixed at 50% of the probability map values. This choice

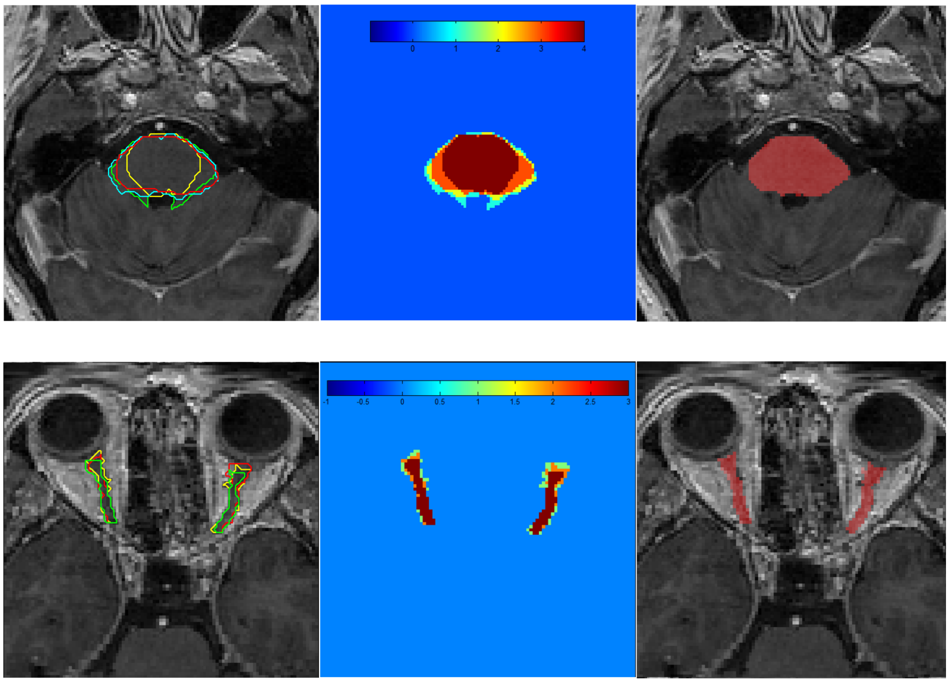


Figure 5.5: Creation of a reference contour example. In the left there row are contours from observers in a 2D axial slice. In the middle row, contours overlapping is shown. Last, in the right, the reference contour is created by majority voting rule from the overlapping map.

for thresholds corresponds to the values proposed by the work of Biancardi et al. [185]. In their work threshold values of 50% and 75% tended to produce consistently large or small estimates, respectively. In figure 5.5, the generation of the reference standard for the brainstem and the optic nerves in our study is shown. When only one manual contour was available, it was directly employed as reference standard (i.e. for eyes and lenses).

Due to differences between observers, generated reference could not always be satisfactory and considered as corrupted data, particularly if they are employed for learning. To ensure this not to happen, an external expert reviewed the generated reference contours and performed small modifications, if needed.

5.4 Leave-One-Out-Cross-Validation

Typical validation techniques to evaluate the performance of a classifier comprises the separation of the available dataset into two independent groups: training and testing group. Accordingly, the training group is used to train the classifier, whereas the testing group is employed to evaluate its performance. Nevertheless, there could be some cases where the availability of images is limited and such division cannot be conducted if a relevant evaluation is envisaged. Such is the case of the dataset employed in this thesis. In these situations, a strategy called Leave-one-out cross-validation (LOOCV) is usually employed. LOOCV is closely related to the validation set approach explained in section 4.4.2.4.1. The difference lies in the attempt of addressing the drawbacks of the later. Like the k-fold CV approach, LOOCV involves splitting the training set into two parts. However, instead of creating k subsets of comparable size, a single observation (x_1, y_1) is used for the validation set, and the remaining observations $(x_2, y_2), ..., (x_n, y_n)$ are used to carry out the training. The learning method is fit on the $n - 1$ training observations, and a prediction $\hat{y_1}$ is made for the excluded observation, using its value x_1. The procedure can be repeated by employing the observation (x_1, y_1) as validation set, and training the statistical learning process on the $n-1$ remaining observations, $(x_1, y_1), (x_3, y_3), ..., (x_n, y_n)$.

This variation of CV can be seen as the k-fold cross-validation where k is equal to the number of samples in the sample set. There is no need to generate random permutations for leave-one-out cross-validation and repeat the process, because the training and validation datasets for each of the folds are always the same, and therefore the result of the accuracy estimation is determined.

One of the major advantages of using LOOCV over the validation set approach is that it has less bias. If we remember, in the validation set method, the training set is commonly half size of the entire dataset. On the other hand, when using LOOCV, the statistical learning approach is repeatedly fitted using training sets which contain $n - 1$ observations. This helps to the LOOCV strategy to have a tendency of not overestimating the test error rate as much as the validation set approach does. Second, since there is no randomness in the training and validation groups, performing LOOCV multiple times will necessarily produce the same outcomes. Contrary to the validation set approach, where results will be different due to the randomness when creating the training and validation sets.

Unlike in Section 4.4.2.4.1, where the partitioning of the data was done by grouping single instances randomly selected from all the patients into the different subsets, in this stage a patient is considered as a sample. That is, during model selection each observation represented a voxel and its features, whereas in classification an observation is assumed to be a patient.

5.5 Evaluation metrics

Medical image segmentation is an important processing step in medical image analysis. Segmentation methods with high precision, high reproducibility and low bias are a main goal in radiotherapy because they directly impact the results. Accurately recognizing some patterns is of great value when segmenting medical images. Consequently, assessing the accuracy and the quality of segmentation algorithm is of great importance. There are different quality aspects in medical image segmentation according to which types of segmentation errors can be defined. Evaluation metrics are expected to indicate some or all of theses errors, depending on the data and on the segmentation task. Requirements of medical segmentation evaluation were categorized by [186] into accuracy, precision as a measure of repeatability and the efficiency. The accuracy category represents the degree of agreement of the segmentation with respect to the reference contours. Under this category, two quality aspects were mentioned, namely the contour, or delineation of the boundary, and the size, or volume of segmented object.

As pointed out by [6], evaluation methods have lacked consensus as to comparison metrics. Since each metric yields different information, their choice is important and must be considered in the appropriate context. Although volume-based metrics, such as Dice Similarity Coefficient (DSC) [187], have been broadly used to compare volume similarities, they are fairly insensitive to edge differences when those differences have a small impact on the overall volume. Therefore, two segmentations with high degree of spatial overlapping may exhibit clinically relevant differences at the edges. As a consequence distance-based metrics, such as Hausdorff distances, are also used to evaluate segmentation results.

Let us now introduce some metric definitions that will be used throughout this chapter. Let a medical volume be represented by a point set $X = \{x_1, ..., x_n\}$, where x_n represent the voxel n. Let denote $|X|$ as $w \times h \times d = n$, where w, h and d are the width, height and depth on the grid where the volume is defined. To facilitate the understanding of following sections, let assume that we only deal with segmentations that have two classes: the class or structure of interest and the background. To refer to the class of interest we will use the number 1, while we will employ the number 2 to refer to the background.

Let denote the volume used as reference V_{ref}, which is represented by the partition $\{V_{ref}^1, V_{gt}^2\}$ of X. The assignment function $f_{ref}^i(x)$ therefore provides the membership of the structure x in the subset S_{ref}^i, where:

$$f_{ref}^{i}(x) = \begin{cases} 1 & \text{if } x \in V_{ref}^{i} \\ 0 & \text{if } x \notin V_{ref}^{i} \end{cases} \tag{5.1}$$

On the other hand, let refer to V_a as the automatic segmentation to be evaluated, which is represented by $\{V_a^1, V_a^2\}$ of X. Similarly to the case of the reference volume, the assignment function $f_a^i(x)$ provides the membership of x in the class S_a^i, which is analogously defined.

5.5.1 Spatial overlap based metrics

Spatial overlap based metrics can be derived from the four basic cardinalities of the so-called confusion matrix: *the true positives (TP), the false positives (FP), the true negatives (TN) and the false negatives (FN).*

5.5.1.1 Basic cardinalities

Let S_a and S_b be two segmentations, the confusion matrix represents the four common cardinalities which reflect the overlap between them: *TP, FP, TN and FN*. For each pair of subsets $i \in S_a$ and $j \in S_b$, the cardinalities provides the sum of agreement m_{ij} between them as follows:

$$m_{ij} = \sum_{n=1}^{|X|} f_{ref}^{i}(x_n) f_a^{i}(x_n) \tag{5.2}$$

where $TP = m_{11}$, $FP = m_{10}$, $FN = m_{01}$ and $TN = m_{00}$. To simplify its definition we can refer to TP as the positive samples that were correctly labeled by the classifier, while TN denote the negative samples correctly labeled. On the other hand, FP represent the negative samples incorrectly classified, i.e. erroneously indicates the presence of a condition, such as a disease, when in reality it is not, for example. Last, and contrary to FP, FN represents an error indicating no presence of a condition when it actually exists. In medical domain, and more generally in binary classification, it is a common practice to directly use these basic cardinalities to assess the performance of a classifier.

5.5.1.2 Dice Similarity Coefficient

The Dice Similarity Coefficient (DSC) has been broadly used in the field of segmentation as a measure of spatial overlapping [187]. As it has been used in the literature, it compares a pair of volumes (binary masks) and provides a similarity index between these two structures. The similarity index or coefficient is defined as the ratio of twice the common area to the sum of the individual areas. Following the nomenclature already introduced, the Dice similarity coefficient is defined as

$$DSC = \frac{2|V^1_{ref} \cap V^1_a|}{|V^1_{ref}| + |V^1_a|} = \frac{2TP}{2TP + FP + FN} \tag{5.3}$$

According to 5.3, DSC values closer to 1 reflect high spatial agreement, while DSC values closer to 0 show poor agreement between the volumes.

5.5.1.3 Sensitivity and specificity

Additionally to volume and distance-based metrics, sensitivity and specificity were also investigated. Sensitivity measures the percentage of actual positives values which are correctly identified whereas specificity measures the percentage of negative values which are correctly identified. To do this, the numbers of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) voxels were determined. These two metrics are defined as follows:

$$Sensitivity = Recall = TPR = \frac{TP}{TP + FN} \tag{5.4}$$

$$Specificity = TNR = TPR = \frac{TN}{TN + FP} \tag{5.5}$$

The sensitivity might be equal to 1 for a poor segmentation much bigger than the ground truth. On the other hand, the specificity, is therefore the necessary counterpart of the sensitivity, but it might tend to 1 for a very poor segmentation that does not detect the object of interest at all. Consequently, a good segmentation system should have high sensitivity and specificity values. It is worth to notice that both measures are very sensitive to the size of the structure of interest. Thus, they penalize errors in small segments more than in large segments [186].

Receiver operating characteristic (ROC) analysis is usually employed to analyze classifiers performance. In this evaluation, curves defining the relation between sensitivity and (1 - specificity) are plotted. If the ROC analysis is considered from a radiotherapy point of view, FN and FP voxels must be taken into consideration when analyzing the segmentation performance. While FN voxels might lead to overirradiation of OARs voxels, FP voxels could result in a possible underirradiation of target volume voxels. Thus, the higher the sensitivity, the lower risk of overirradiation of normal tissue and the higher the specificity, the lower the risk of underirradiation of tumor tissue. Following the suggestion of [188], instead of employing ROC curves to evaluate performance of a given classifier, the ROC space is used. The ROC space can be divided into four sub-spaces. This sub-division scheme is shown in Figure 5.6. Thus, results spread over the left-top sub-space indicate acceptable contours, with the OAR spared and the PTV covered. Results lying on the right-top sub-space present a high-risk, since the OAR may be spared but with PTV not covered. Poor contours are considered when they ROC representation are present on the left-bottom sub-space. There, although the PTV is covered, it is considered that the OAR is not spared. And last, the right-bottom side of the ROC subdivision contains the unacceptable contours, with OARs not spared and PTV not covered.

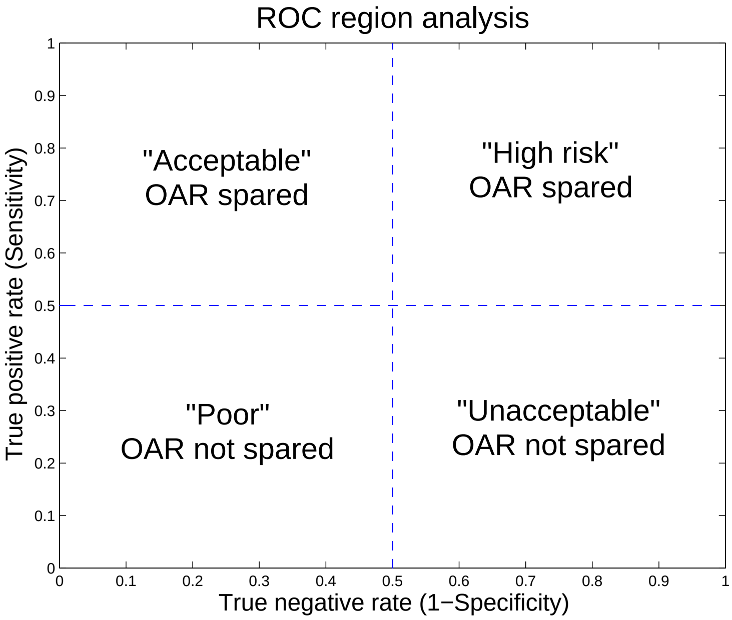


Figure 5.6: ROC space sub-division to evaluate our classifier performance.

5.5.2 Volume based metrics

It is important to note that, although the concept of OAR is purely oncological or anatomical, a representation of these volumes is used in the planning process. Therefore, the defined volume of a critical structure plays a crucial role in the dose distribution planned. As can be seen in the table 2.1, where the dose limits for the OARs in both radiotherapy and radio-surgery are defined, especially in the case of radio-surgery, variations in the volume may lead to variations in the planned dose.

Consequently, volume based metrics are of significant importance when generating contours to be used in the RTP. As its name indicates, volume based metrics are measures that consider the volumes of the segmentations to indicate similarity. To measure volume differences between manual and automatic contours with the reference, we consider the the following formula

$$\Delta V(\%) = \frac{V_a^1 - V_{ref}^1}{V_{ref}^1} * 100 \tag{5.6}$$

which will be referred to as *relative Volume Differences* (rVD). An important point to note in this metric is that, while the absolute value of $\Delta V(\%)$ will be used to plot rVD values and to compute mean values, values directly obtained from eq. 5.6 (either negative or positive) will be used in the statistical analysis. The reason to employ absolute values of rVD to compute means is to evaluate total relative differences between contours. If, for example, we consider two contours that differ from the reference standard in -10 and 10%, the mean will be 0 if negative values are also taken into account. However, both contours will have a difference with respect to the reference of 10%, independently of the sign, leading a mean deviation of 10%.

5.5.3 Spatial distance based metrics

To tackle with edge dissimilarities that have a small impact on the overall segmented volume, volume-based metric are not sufficient. If a given segmentation is planned to be used in RTP, an analysis on shape fidelity of the segmentation outline is highly recommended. Any underinclusion on the OAR delineation might lead to a part of the healthy tissue exposed to radiation. Spatial distance based metrics have been also widely employed in the literature to evaluate image segmentations as dissimilarity measures. They are strongly recommended when the segmentation overall accuracy is crucial, as in the case of its inclusion in the RTP. Therefore, a surface distance measure (Hausdorff distance [189]) was also used to evaluate the segmentation results.

5.5.3.1 Hausdorff Distance

The Hausdorff Distance is a mathematical construct to measure the ”closeness” of two sets of points that are subsets of a metric space. It represents the ”maximum distance of a set to the nearest point in the other set”. More formally, Hausdorff distance from the finite point set $X = \{x_1, ..., x_p\}$ to the finite point set $Y = \{y_1, ..., y_p\}$ is a *maximin* function, defined as

$$H(X,Y) = max(h(X,Y), h(Y,X)) \tag{5.7}$$

where

$$h(X,Y) = \max_{x \in X} \min_{y \in Y} \|x - y\| \tag{5.8}$$

and $\| \cdot \|$ is some underlying norm on the points of X and Y, such as L_2 or the Euclidean Norm.

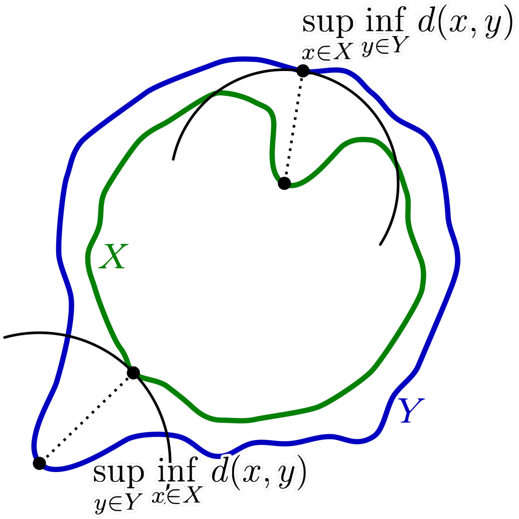

$\sup_{x \in X} \inf_{y \in Y} d(x, y)$
X
Y
$\sup_{y \in Y} \inf_{x \in X} d(x, y)$

Figure 5.7: A schematic figure explaining the concept of Hausdorff distance with two segmentation proposals, X and Y, for a certain structure.

Using Figure 5.7 as example: ROI X and ROI Y are two different segmentation proposals in a single MRI slice under evaluation. Somewhere on the edge of ROI X there is a point, x, that is further away from any point on Y than all other points on X's edge. This point has a minimum distance, l_2, to ROI Y. This is the Hausdorff distance from X to Y. Similarly on the edge of ROI Y there is a point y that is further away from any point on X than all other points on the edge of Y. The minimum distance, l_1, from point y to a point on the edge of X is the Hausdorff distance from Y to X. The maximum of these two values (the longer of the two lines), in this case l_1, is the Hausdorff distance between ROI X and ROI Y in this MRI slice.

Thus, this distance can be used to determine the degree of resemblance between two objects that are superimposed on to another [189].

5.5.4 Efficiency

Efficiency describes the practical viability of the segmentation method. It refers to the practical viability of a segmentation method. Two factors need to be considered to fully characterize efficiency: computational time and the human operator time required to complete segmentation of each study in a routine setting in the application domain. As it was already presented, user interaction is minimized to a simple alignment of that target patient before to send it to the classification process. Therefore, to assess efficiency, the computational time required for algorithm execution should be measured and analyzed.

5.5.4.1 Processing Time

For comparison purposes, segmentation time observed for each physician when manually segmenting the OARs was recorded. The segmentation of each structure was timed individually both for the manual segmentation and for the automatic contours. The total time per patient for each of the methods was then compared as well as the time consumed per structure. For the purpose of our application, we can consider two different times through the whole segmentation process: *features extraction* and *classification* time.

5.6 Statistical analysis

Among different types of inferential statistical tests, analysis of variance (ANOVA) are the most suitable one for the purpose of our evaluation. ANOVA is a parametric method for means comparison of several groups and it tests the significance of group differences between two or more groups. It is important to point out that it only determines that there is a difference between groups, but it does not tell us which is different. In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups.

The first of the techniques encompassed in ANOVA approaches is the one-way ANOVA. It is used to determine whether there are any significant differences between the means of two or more groups. However, one of the assumptions is that samples contained in the groups must be independent, which is not the case in our study.

Nevertheless, one-way repeated measures ANOVA is the equivalent of the one-way ANOVA, but for related -not independent- groups. A repeated measures ANOVA is also referred to as a *within-subjects ANOVA* or *ANOVA for correlated samples*. All these names imply the nature of the repeated measures ANOVA, that of a test to detect any overall differences between related means. One-way repeated measures ANOVA compares how a within-subjects experimental group performs in three or more experimental conditions. This means that it is used when you have a single group on which you have measured something a few times. The analysis compares whether the mean of any of the individual experimental conditions differ significantly from the aggregate mean across the experimental conditions.

Particularly, by employing statistical analysis we aim at demonstrating that differences in volume and surface were significantly different between SVM and our SDAE-based classification system. On the other hand, we also employed statistical analysis between manual annotations and contours generated by our system. In this case, we expect to prove that, although results from some manual observers were better than results provided by our approach, differences were not significantly important.

Chapter 6

Experiments and Results

" There is only one way to avoid criticism: do nothing, say nothing, and be nothing."
Aristotle

This chapter focuses on the experiments that were carried out to achieve the results presented on this work, and how they were implemented. Setting-up of these experiments is detailed in the first section on the chapter. The parameterization of all the values involved in any step of the proposed work-flow are detailed in this section. This includes steps such as generation of probability map or common mask, the composition of the features vector, how features were extracted and choice of SDAE parameters, for example. The second section presents the results that come from the experiments. For comparison purposes, the proposed method is always compared against a classifier based on SVM. Then, manual observers are also taken into account to evaluate the performance of our proposed scheme in clinical settings. The main objective of this section is to demonstrate that our proposed scheme outperforms SVM when classifying OARs in brain cancer, as well as it lies in the variability of the experts. Accordingly, results are subdivided into subsections that details the obtained results. Last section summarizes the results, and a discussion about them is presented. Comparison with other presented methods to segment OARs in brain cancer is presented in this section.

6.1 Experiments set-up

In chapter 4, the theoretical introduction on how parameterization must be done has been introduced. Now, in this section, values obtained through the parameterization employed in all the steps are detailed.

6.1.1 Parametrization

In previous sections, a detailed theoretical explanation of how training and classification has been performed was introduced. There, reasonings about procedures followed to train the learning based systems were detailed in order to support their use. In following sections, parameters used in each of these processes are presented.

6.1.1.1 Probability map and common mask creation

We have previously seen that the first step right after aligning the images contained in the training set is generating a spatial probabilistic distribution map (SPDM) for each of the OARs. To generate the SPDM, aligned manual labels are added into a volume. The resulted image is then smoothed by using a Gaussian filter with a kernel size of 3x3x3. To reduce the number of input samples that contain consistent information, the voxel space was first binarized by setting its values greater than 0.005 to 1, and the others to 0. Then, a dilation operation with a square kernel type of size 3x3x3 was applied over the binary image. Only those voxels that belonged to the inner part of the dilated image were kept to extract the features.

6.1.1.2 Composition of the features vector

As we introduced in 4.3, dissimilarities between characteristics of OARs cause that some of the suggested features are organ dependent, not being suitable for all the organs investigated. Thus, two groups of OARs have been identified: large and/or well-defined organs with no large shape variations, and organs which texture is heterogeneous and/or large shape variations and which localization also presents a high variation. From now onward, they will be referred to as group A, and B, respectively. See table 6.1 for a classification of OARs in both groups.

OARs groups classification	
Group A	Brainstem Eyes Lenses
Group B	Optic nerves Pituitary gland Pituitary stalk Chiasm

Table 6.1: Classification of the OARs in groups A or B.

To demonstrate that including proposed features, for each group, positively impacts on the segmentation performance, different features sets have been evaluated. Thus, the first set for each group is composed by features that have already been proposed in other works. This set will be referred to as *classical* features in all the groups. Several features sets were investigated depending on the OARs group. A complete list describing the composition of features vectors used is presented in table 6.2. Next section presents the details of how features were extracted for each of the groups.

Features set name	Features included	Vector size
Group A		
Classical	Intensity of voxel under examination Intensity of voxel neighborhood (3D) Intensity of 8 voxels along maximum gradient direction Probability voxel value Spherical Coordinates	39
Enhanced	Classical (except 3D voxel neighborhood) Geodesic Distance Transform Map 3D-Local Binary Texture Pattern Gradient value of voxel	19
Group B		
Classical	Intensity of voxel under examination Intensity of voxel neighborhood (3D) Intensity of 8 voxels along maximum gradient direction Probability voxel value Spherical Coordinates	137
Augmented	Classical Gradient Patch in 2D (Horizontal and vertical magnitudes and orientation) Contextual features	276
Textural	Classical Mean Variance Entropy Energy Kurtosis Skewness Wavelet patch decomposition	149
AE-FV	Classical Augmented Textural	288

Table 6.2: Features sets employed for the different groups.

6.1.1.3 Features Extraction

MR T1 sequence was the only image modality used. Intensity information around neighboring region of the voxel under examination was extracted by employing three-dimensional patches in groups A and B. However, patch size in group A was 3x3x3 whilst in group B it was 5x5x5. The reason for this difference is that OARs included in group A present a more homogenized texture than those included in group B. Furthermore, we experimented with both sizes in OARs of group A, and no significant improvement was found. Following the same reasoning, intensity neighborhood properties in OARs of group C were extracted in a patch of size 5. Nevertheless, instead of extracting the information of a three-dimensional vicinity, only the 2D space was taking into account. Therefore, vector sizes for classical features resulted to be of 39, 137 and 34, for groups A,B and C, respectively.

Suggested features to segment OARs of group A include the use of a geodesic distance transform map (GDTM)(section 4.3.3), the proposed 3D-Local binary texture pattern (3D-LBTP) (section 4.3.4) and the gradient value of the voxel under examination. The GDTM was generated by employing the 3D input image. To calculate the value of the GDTM at each voxel, we used a patch of size 3x3x3 and λ was set to 0.75. As detailed in section 4.3.4, 6 voxels around the central voxel and ratios equal to 1 and 2 were employed to capture the neighborhood appearance. In total, 4 values were extracted for this feature: 1 texture and 1 binary values at each ratio. This led to a features set composed by 19 features.

Features proposed to segment OARs belonging to group B are divided into three groups: *augmented, textural and augmented-enhanced features vectors*. In addition to specific features for each group, they include features described for the classical features set. Gradient information was extracted on a two-dimensional patch of size 5x5 around each voxel for each of the gradient properties (horizontal and vertical gradient values, as well as gradient orientation). Thus, 75 gradient values were obtained for each voxel. In addition to gradient, contextual features were also included in this set. As in the work of [168] regions of size 3x3x1 voxels were sampled around the voxel under examination by radiation from it at every 45°, and at four different radius: 4,8,16 and 32. By combining the continuous and the binary value at each sampled patch, this led to a total of 64 contextual features for each voxel. Textural features set comprises features related with texture. To compute first-order textural features, patches of size 3x3x3 were extracted around each voxel. Additionally, for the skewness, kurtosis and entropy, an additional patch of size 5x5x5 was also employed, leading to a two values of these features for each voxel (one value per patch). In addition to these patch sizes, other different patches configurations were investigated. Particularly, patches of size 7, 9 and 11 were included in the features vector. However, their inclusion did not lead to significant performance improvement, but it considerably increased the computation time to extract the features. Therefore, they have not been included in our evaluation. Regarding the use of wavelet-based features, first to fourth order high-pass components from discrete wavelet decomposition were employed. Total number of features used in each features set is shown in table 6.2. And last, the features set named augmented-enhanced features vector encompasses all the sets previously presented for OARs of group B. Therefore, sizes for each of the features sets are as follows: 137 for the *classical* set, 276 for the *augmented* set, 149 for the *textural* set and 288 for the proposed $AE\text{-}FV$ set.

6.1.1.4 Features scaling

Figure 6.1 shows the distribution of some features representing optic chiasm and non optic chiasm samples for one patient. For the purpose of visualization, only few features have been selected. The idea is to show that features included in the vector incorporate additional discriminative information for the segmentation. To avoid features with greater values dominating the classification, the features vector was normalized before training or testing. Except for the BRIEF descriptor features, all the rest were normalized in the range of $[-1, 1]$. The same scaling factors applied during training are employed in the classification. To demonstrate that normalizing the features values does not affect to their discriminative power, the distribution of normalized features is plotted in the lower row of figure 6.1.

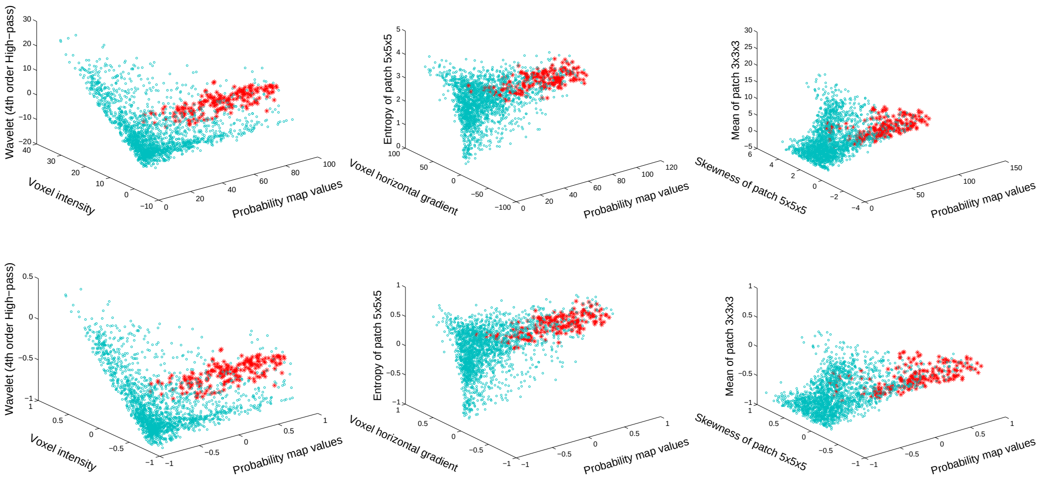


Figure 6.1: Scatter plots of samples from the same subject showing different features sets representations for the optic chiasm. Red crosses and blue circles indicate optic chiasm and non optic chiasm samples, respectively. While the upper row plots samples non-normalized, the row on the bottom represent the distribution of normalized features.

6.1.1.5 Parameter setting for SVM

The two parameters that can be tuned in the RBF kernel and which depend on the input data are: C and γ. A coarse grid search, followed by a finer search was performed to find the best combination of both parameters. For example, for the brainstem case it was found from this search that best values for C and γ were approximately 6 and 5.5, respectively, with an accuracy close to 97% and a precision nearly of 95% (Fig. 6.2). These values for C and γ were kept for the training and classification in all the features set.

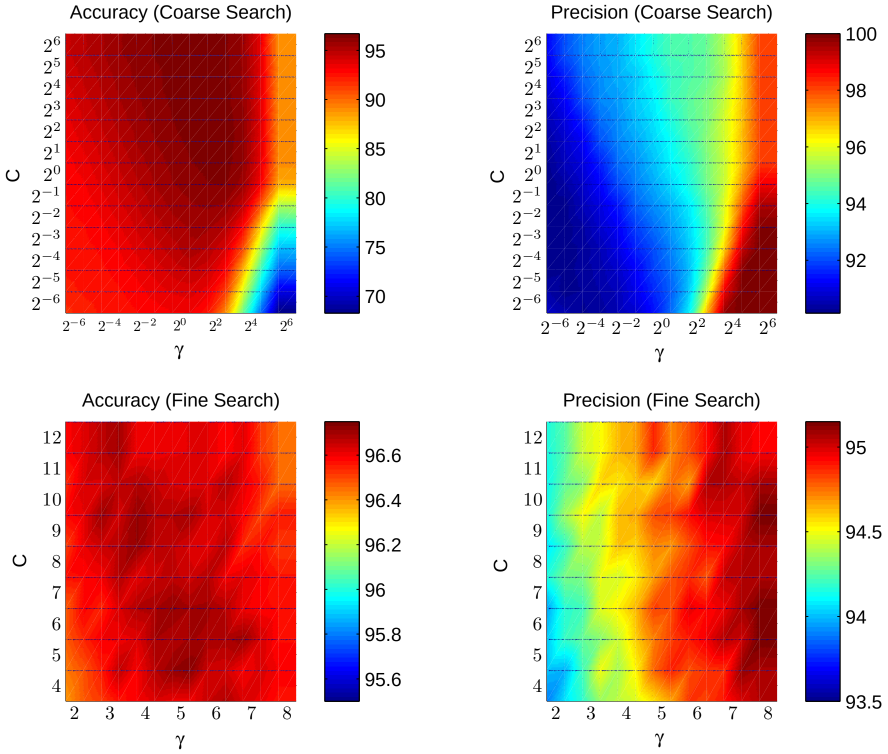


Figure 6.2: Parameter setting for SVM with different C and lambda values for the brainstem case.

6.1.1.6 Parameter setting for SDAE

The deep network used in the proposed classification scheme was formed by stacking DAEs (Fig. 6.3). Weights between layers of the network are initially learned via the unsupervised pre-training step. Once all the weights of the network are unsupervisedly computed, a supervised refinement is carried out by using the labeled classes, and final values of the network' weights are updated (Sec. 4.2.7).

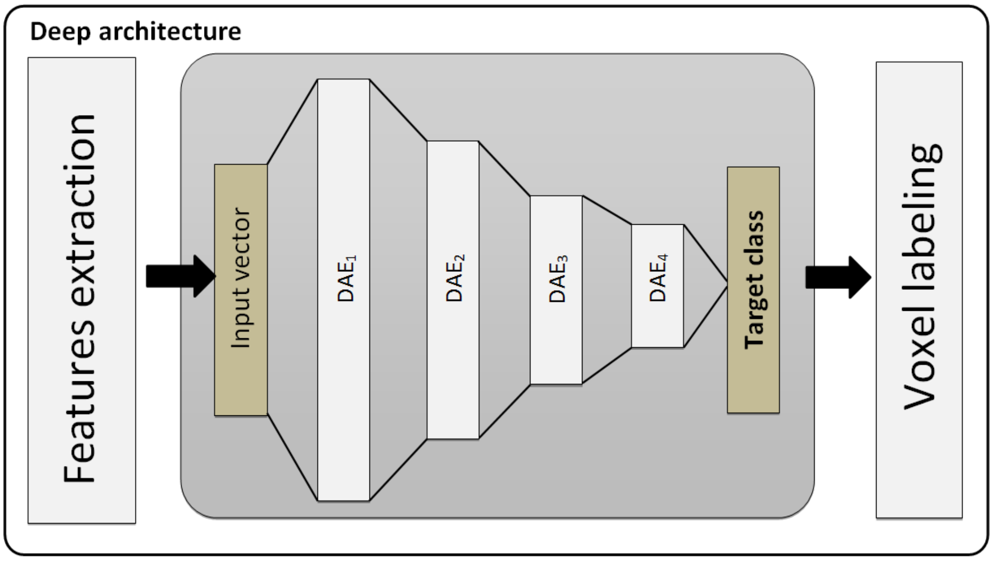


Figure 6.3: Deep network architecture constructed by stacking denoising autoencoders in the proposed approach.

The stack of DAEs forms the intermediate layers of the deep network (See Figure 6.3). Nevertheless, defining the number of hidden layers, as well as their size, is not an easy task. Training was run multiple times with different configurations of the deep architecture to find a proper combination of parameters. As introduced in section 4.4.2.4.2, the strategy followed to find a network configuration is based on the error convergence during training. Curve plotted in figure 6.4 shows the progression of this error for several network configurations. With this procedure we obtained two optimal network configurations, which depends on the number of elements composing the features vector (Table 6.2). Architecture of networks aiming at segmenting OARs of group A was composed by 4 hidden layers, with 100, 50, 25 and 10 units, from input to output, respectively. On the other hand, for OARs of groups B, the network structured was composed by 4 hidden layers, with 400, 200, 100 and 50 units, from input to output, respectively. The learned representation of the input had therefore a dimensionality of 10 for the structures of group A and 50 for structures of group B.

Since our network is composed by 4 hidden layers, during the unsupervised pre-training, the weights vectors $\{W_1, W_2, W_3, W_4\}$ were initially learned. Denoising corruption level for the DAEs was set to 0.5, since a value of 50% of noise level has already been proved to perform well in other problems [159]. Following the same architecture than in the unsupervised pre-training, four hidden layers of DAEs were used for the fine-tuning step, with the same number of units than before. At the end of the last layer of DAEs a logistic regression layer is used as output with the sigmoid function as activation function.

SDAE parameters	
Network Structure	Group A 100 - 50 - 25 - 10 Group B 400- 200 - 100 - 50
Number of classes	2
Corruption Level	0.5
Batch sizes (per layer)	200 - 500 - 1000 - 2000
Number of epochs	500
Learning rate	0.1
Activation function (Unsupervised learning)	Sigmoid
Activation function (Supervised learning)	Sigmoid
Output	Logistic

Table 6.3: Summary of employed SDAE parameters.

Mini-batch learning was followed during both unsupervised pre-training of DAEs and supervised fine-tuning of the entire network. Batch sizes were set in both configurations to 200,500,1000 and 2000, from the top to the bottom layers, respectively. Table 6.3 summarizes parameter values employed in the proposed SDAE.

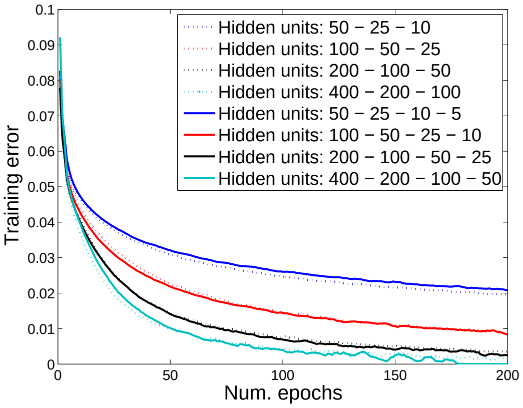


Figure 6.4: Evolution of batch errors during training for different configurations of the deep architecture. Epochs refers to the number of passes through the data.

6.1.1.7 Number of features

It is very common in practice to have a training set with unbalanced number of positives and negatives samples. Not taking the proper balance between them might lead to unsatisfactory results. Since the performance of the classifier depends on the available data, there exist no rule to define the best balance between positive and negative samples in a training set. Therefore, we evaluated the impact of different unbalanced training sets on the DSC. Figure 6.5 plots the evolution of the DSC in relation with the proportion of positive and negative samples for a patient in a given OAR. Since the number of negative samples was much higher than positive samples, we employed all the positive samples and increased by steps the number of negative samples. We observed that, in general, DSC increased up to having a number of negative samples equal to 32 times the number of positive ones. Increasing the number of negative samples beyond 32 did not significantly improve DSC values. The reason of this behavior can be attributed to the amount of data often required from deep learning methods to learn input representations. Some structures in our experiment were composed by an average amount of voxels ranging from 80 to 785. This is the case, for instance, of the chiasm, which mean volume was composed by 235 voxels. Due to the limited available dataset, the training set for the chiasm in the balanced case was composed by nearly 6580 voxels (235 × 14 patients × 2). This number of samples showed to be insufficient for providing the best volume similarity performance in our experiment. Particularly, a low amount of available samples for training makes the situation even worst in cases presenting a large variability between samples, such as for the optic nerves. Regardless of type or sample, i.e. either negative or positive, by adding more samples on the training set increased the volume similarity performance. Thus, for training purposes, the number of negatives samples was 30 times the number of positives samples, when that amount of negative samples were available. Otherwise, all the samples were taken into consideration to train the classifier.

6.1.2 Leave-one-out-cross-validation

As explained in section 5.4, we employ this strategy to evaluate our method. This technique consists in leaving one of the patients of the dataset out, and train the classifier by using the remaining patients. This process is repeated as many times as available patients we have. Thus, taking into account that our dataset is composed by 15 patients, we will use 14 patients for training and 1 for classification, which will be repeated 15 times, leaving one different patient in each iteration.

6.2 Results

Since SVM has proven to be a state-of-the-art classifier, we use it in this thesis for comparison purposes. To demonstrate that employing a deep network

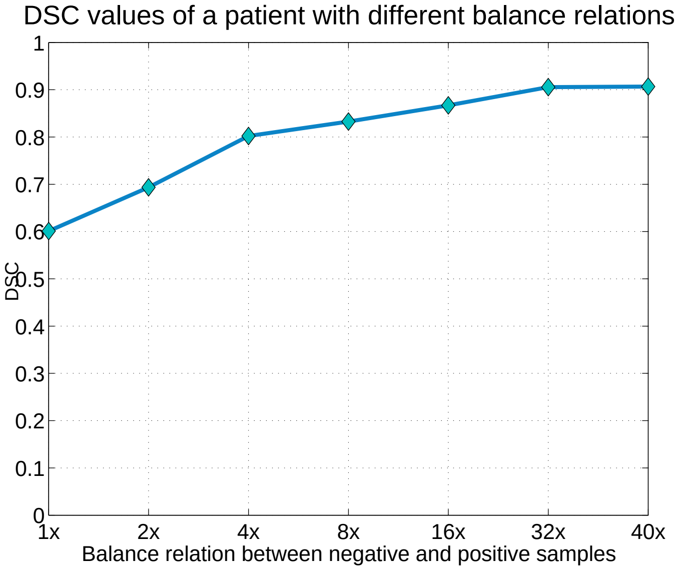


Figure 6.5: DSC values for a given patient with different balance relations between the number of positives and negative samples used to train the classifier.

scheme to classify OARs can outperform SVM, different configurations were evaluated. Changes on configurations comprise: i) the use of either SVM or SDAE for classification and ii) the use of one of the features sets described in Section 6.1.1.2. Accordingly, the first configuration will always be composed by SVM and classical features, which will be referred to as SVM_1. Next, classical features will be employed in the SDAE based system, which leads to the configuration known as SDAE_1. Depending on the OARs group, several configurations will be evaluated (Table 6.2). Accordingly, configurations will be referred to as SDAE_n, where n denotes the features group used. Finally, SVM will be employed with the last features set of each configuration, i.e. proposed set, leading to the SVM_2 or SVM_{AE-FV} set for organs from group A or B, respectively.

Structures considered as OARs in the present work differ between them in texture and/or shape appearance. Nevertheless, as explained in section 4.3, despite these differences, there are some structures that present a sort of homogeneity in texture and variation in shape and location is less strong than in others. Therefore, OARs are classed into two main groups, A and B. As a remainder, the group A is composed by the brainstem, eyes and lenses. On the other hand, optic nerves, optic chiasm, pituitary gland and pituitary stalk are considered to belong to the group B. Because of proposed features vary between group A and B, results for both groups are presented separately.

Additionally, number of patients and manual contours available to evaluate the performance of the proposed approach over different OARs was different (See section 5.3 to see the dataset composition). Hence, results that show comparisons between manual and automatic contours are presented when available.

6.2.1 OARs group A

This section presents results of the automatic approaches to segment OARs that belong to group A. For evaluation purposes, in those organs separately present in both left and right brain sides, each of the sides are individually analyzed. With regards to features sets employed, we refer to the table 6.2, where different groups of features were presented. Following this definition and as previously explained, the deep learning scheme that employs classical features will be referred to as SDAE_1, whilst the one employing the proposed set will be referred to as SDAE_2 in this section. In addition, the setting employing SVM and classical features will be referred to as SVM_1, while the configuration employing the proposed features will be referred to as SVM_2.

6.2.1.1 Comparison with respect to the reference standard

Performance of the four automatic configurations with respect to the reference standard is evaluated in this section. The objective is to quantitatively demonstrate that our proposed learning scheme outperforms the SVM settings, as well as the SDAE scheme configured with classical features.

Dice Similarity Coefficients. Dice similarity coefficients obtained by the automatic segmentations of the OARs of group A are plotted in Figure 6.6. Box plots are grouped for each OAR. Inside each group, results for the reference SVM_1, the SVM scheme employing proposed features (SVM_2), the SDAE setting with classical features (SDAE_1) and our proposed system (SDAE_2) are displayed. Median values for each group were taken to compute the 50% percentile of the distribution, q_{50}. To calculate the first and third quartile, i.e q_{25} and q_{75}, median values of elements lower and higher than q_{50} were respectively employed. Then, the Interquartile range (IQR) was equal to q_{75} - q_{25}. The lower and upper inner fences were estimated taking 1.5×IQR from the quartile (the "inner fence") rather than the max or min. Last, outliers were those values that were either 1.5×IQR or more above the third quartile or 1.5×IQR or more below the first quartile. Looking across all structures, segmentations produced by SVM_1 system achieved the lowest results in comparison with the other three settings. While it reported a mean DSC of 0.77 (± 0.05) over all the structures, a mean value of 0.79 (± 0.04) was achieved by the SVM_2 set-

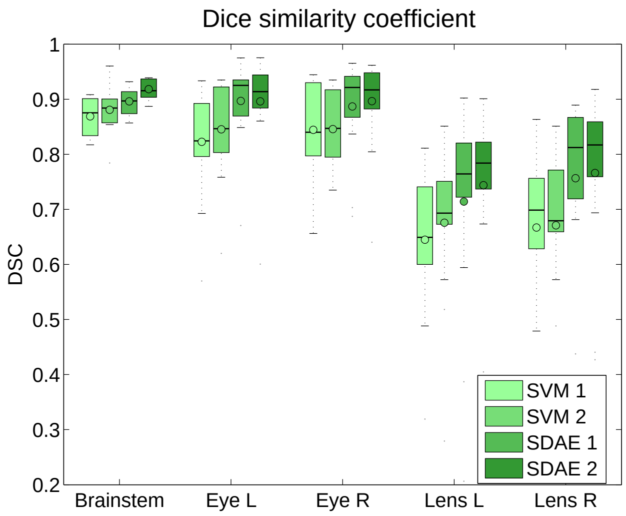


Figure 6.6: Segmentation DSC results for the automatic contours with different settings for organs of group A.

ting. Schemes based on deep learning, i.e. $SDAE_1$ and $SDAE_2$, obtained mean DSC values of 0.83 (± 0.05) and 0.85 (± 0.05), respectively. Decomposing into single structures, it can be observed that solely by employing SDAE in the classification scheme instead of SVM, segmentation performance improved in all the structures, as well as variability was reduced. If, in addition, the proposed features are fed into the classifier, performance still improved in most of the OARs, particularly in SDAE frameworks. Specifically for the proposed configuration, $SDAE_2$, whereas the mean DSC for the brainstem was greater than 0.9 (0.92 ± 0.02), it was close to 0.9 for both eyes. For both lenses, however, mean DSC was nearly 0.75. Furthermore, the overall minimum DSC in large structures was typically above 0.85. For small organs, this minimum value was just below 0.7.

Automatic segmentations presented small but significant ($p < 0.05$) differences across machine and deep learning environments when conducting a within-subjects ANOVA test on the DSC of all the groups. The small p-value indicated that at least one method significantly differed from the others. Paired repeated measures ANOVAs (Table 6.4) shows the p-values obtained when comparing results between only two groups. This table pointed out that differences were particularly notorious on the scheme employing SVM combined with classical features as classifier (first row of each group), which values were lower than 0.05. Regarding the inclusion of proposed features in the deep learning scheme, with exception of the brainstem case ($p = 0.0181$), no significant differences were found on the DSC between the groups using SDAE as classifier.

Paired ANOVA (DSC)								
	Left Eye				Right Eye			
	SVM$_1$	SVM$_2$	SDAE$_1$	SDAE$_2$	SVM$_1$	SVM$_2$	SDAE$_1$	SDAE$_2$
SVM$_1$	1	0.3345	0.0265	0.0396	1	0.9640	0.1094	0.1955
SVM$_2$	-	1	0.1626	0.2061	-	1	0.0773	0.1574
SDAE$_1$	-	-	1	0.9818	-	-	1	0.7560
SDAE$_2$	-	-	-	1	-	-	-	1
	Left Lens				Right Lens			
	SVM$_1$	SVM$_2$	SDAE$_1$	SDAE$_2$	SVM$_1$	SVM$_2$	SDAE$_1$	SDAE$_2$
SVM$_1$	1	0.5351	0.2543	0.0736	1	0.9402	0.1148	0.0748
SVM$_2$	-	1	0.5315	0.2210	-	1	0.1277	0.0837
SDAE$_1$	-	-	1	0.6442	-	-	1	0.8636
SDAE$_2$	-	-	-	1	-	-	-	1

Brainstem				
	SVM$_1$	SVM$_2$	SDAE$_1$	SDAE$_2$
SVM$_1$	1	0.5173	0.0275	5.7583x10^{-5}
SVM$_2$	-	1	0.3872	0.0323
SDAE$_1$	-	-	1	0.0181
SDAE$_2$	-	-	-	1

Table 6.4: Paired ANOVA tests for the DSC between the automatic approaches to segment OARs from group A.

Hausdoff distances. Figure 6.7 presents the values of Hausdorff distances obtained for the four configurations. SVM based systems achieved the highest overall mean HD values among the four groups, with values of 5.96 (± 1.11) and 5.23 mm (± 1.02) for SVM$_1$ and SVM$_2$, respectively. On the other hand, these values decreased when employing SDAE as classifier, with mean HD of 4.29 (± 1.09) and 4.07 mm(± 0.98), for SDAE$_1$ and SDAE$_2$, respectively. Having a look to the HD distributions on individual organs on figure 6.7, it can be observed that both SVM settings achieved the highest mean HD values across all the OARs. Although the addition of proposed features into the SVM framework improved mean HD values, it was not sufficient to outperform SDAE based classifiers. Concerning the use of SDAE, mean HD achieved by both settings were very similar when segmenting both eyes and lenses. Across these structures, mean HD values for SDAE$_1$ were 5.35 (± 3.35), 5.29 (± 2.58), 2.01 (± 0.73) and 2.27 mm (± 1.04) for left and right eye, and left and right lens, respectively. When employing proposing features, mean values were: 5.10 (± 2.05), 5.21 (± 3.06), 2.06 (± 0.76) and 2.15 mm (± 0.95), respectively. However, the addition of proposed features decreased values of Hausdorff distances with respect to the setting that employed classical features (SDAE$_1$) when segmenting the brainstem. While mean HD achieved by the SDAE$_1$ scheme was reported to be 6.54 mm (± 2.17 mm), mean HD had a value of 5.87 mm, with a lower standard deviation (0.99 mm), if we employed proposed features instead. Results also reported that for large organs, overall maximum distances were around 10-14 mm when employing SVM in the classification scheme. On the other hand, these maximum values decreased to almost half in SDAE settings, not typically exceeding the barrier

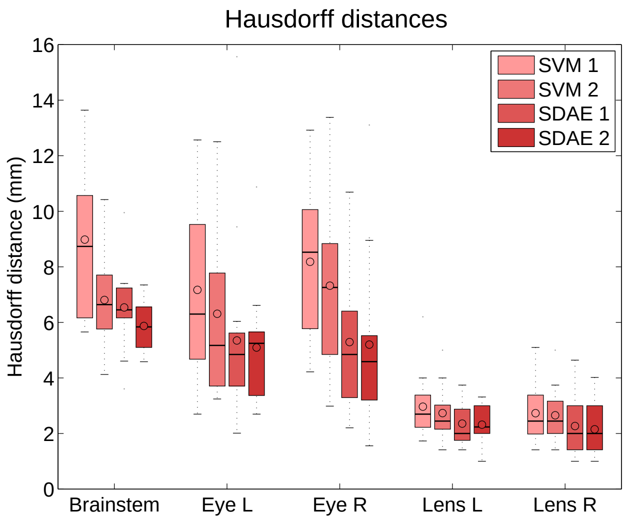


Figure 6.7: HD results for the automatic contours with different settings for organs of group A.

of 8 mm. For the lenses, however, maximum HD values were below 5.5 mm in all configurations.

Paired ANOVA (HD)								
	Left Eye				Right Eye			
	SVM$_1$	SVM$_2$	SDAE$_1$	SDAE$_2$	SVM$_1$	SVM$_2$	SDAE$_1$	SDAE$_2$
SVM$_1$	1	0.4673	0.1484	0.0408	1	0.4227	0.0067	0.0098
SVM$_2$	-	1	0.4175	0.2093	-	1	0.0545	0.0634
SDAE$_1$	-	-	1	0.8042	-	-	1	0.9286
SDAE$_2$	-	-	-	1	-	-	-	1
	Left Lens				Right Lens			
	SVM$_1$	SVM$_2$	SDAE$_1$	SDAE$_2$	SVM$_1$	SVM$_2$	SDAE$_1$	SDAE$_2$
SVM$_1$	1	0.5541	0.0926	0.0759	1	0.8392	0.2463	0.1329
SVM$_2$	-	1	0.2356	0.1953	-	1	0.2922	0.1525
SDAE$_1$	-	-	1	0.8927	-	-	1	0.7487
SDAE$_2$	-	-	-	1	-	-	-	1

Brainstem				
	SVM$_1$	SVM$_2$	SDAE$_1$	SDAE$_2$
SVM$_1$	1	0.0307	0.0158	$6.3953\text{x}10^{-4}$
SVM$_2$	-	1	0.7379	0.1440
SDAE$_1$	-	-	1	0.2885
SDAE$_2$	-	-	-	1

Table 6.5: Paired ANOVA tests for the Hausdorff distances between the automatic approaches to segment OARs from group A.

The ANOVA test demonstrated that there existed also differences between automatic segmentations in relation to Hausdorff distances ($p < 0.05$). As in the case of Dice similarities, significant differences mainly come from the SVM$_1$ setting, particularly if it is compared with SDAE groups (Table 6.5). Although results plotted on Figure 6.7 shows that including the proposed features on the classification scheme slightly decreased the values of HD, the

ANOVA analysis indicates that no significant differences between ($SDAE_1$) and ($SDAE_2$) existed with regards to HD. With exception of the brainstem, where differences in the mean of HD was larger, we can say that improvement is therefore marginal in terms of surface difference.

Relative Volume differences. Figure 6.8 plots relative volume differences (rVD) distributions of the four automatic schemes for each of the organs from group A. Schemes employing SVM as classifier presented the largest volume differences for all the structures. Such differences in volume often doubled the value of differences obtained by SDAE classifiers. For example, while volumes generated by SVM_1 and SVM_2 when segmenting the lenses were sometimes around 100-120% larger than the reference standard, these differences were reduced to 50-55% when employing $SDAE_1$ and $SDAE_2$. Analyzing structures individually, mean relative volume differences obtained by SVM_1, in absolute values, were: 15.43 (± 8.40), 36.77 (± 28.93), 26.33 (± 15.03), 54.60 (± 39.52) and 43.74% (± 30.95) for the brainstem, left eye, right eye, left lens and right lens, respectively. When adding the proposed features into the SVM-based scheme these values became: 7.76 (± 4.99), 14.54 (± 8.51), 20.04 (± 21.02), 55.32 (± 38.45) and 42.79% (± 28.34). In the same order, $SDAE_1$ obtained the following relative volume differences: 3.97 (± 2.03), 8.17 (± 6.09), 10.72 (± 6.15), 29.52 (± 26.25) and 20.12% (± 12.52). At last, reported differences for the proposed scheme ($SDAE_2$) were: 3.01 (± 1.23), 10.02 (± 6.01), 10.07 (± 7.74), 28.05 (± 24.14) and 17.57% (± 11.69).

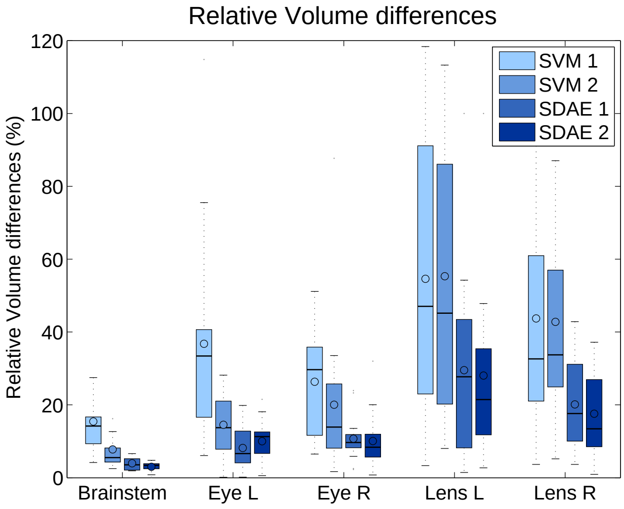


Figure 6.8: Vol diff results for the automatic contours with different settings for organs of group A.

Differences between automatic segmentations in terms of volume were sta-

Paired ANOVA (Vol Diff)								
	Left Eye				**Right Eye**			
	SVM_1	**SVM_2**	**$SDAE_1$**	**$SDAE_2$**	**SVM_1**	**SVM_2**	**$SDAE_1$**	**$SDAE_2$**
SVM_1	1	0.0081	$8.2509 \text{x} 10^{-4}$	0.0015	1	0.1956	$8.6539 \text{x} 10^{-4}$	$1.1963 \text{x} 10^{-4}$
SVM_2	-	1	0.0255	0.0927	-	1	0.1106	0.0788
$SDAE_1$	-	-	1	0.4121	-	-	1	0.3521
$SDAE_2$	-	-	-	1	-	-	-	1
	Left Lens				**Right Lens**			
	SVM_1	**SVM_2**	**$SDAE_1$**	**$SDAE_2$**	**SVM_1**	**SVM_2**	**$SDAE_1$**	**$SDAE_2$**
SVM_1	1	0.9604	$4.5912 \text{x} 10^{-6}$	$2.6051 \text{x} 10^{-6}$	1	0.9312	$4.6127 \text{x} 10^{-6}$	$7.3127 \text{x} 10^{-6}$
SVM_2	-	1	$3.1125 \text{x} 10^{-6}$	$1.1722 \text{x} 10^{-6}$	-	1	$4.4581 \text{x} 10^{-6}$	$7.0858 \text{x} 10^{-6}$
$SDAE_1$	-	-	1	0.6442	-	-	1	0.5680
$SDAE_2$	-	-	-	1	-	-	-	1
		Brainstem						
			SVM_1	**SVM_2**	**$SDAE_1$**	**$SDAE_2$**		
		SVM_1	1	0.0051	0.2013	0.0233		
		SVM_2	-	1	0.0018	0.0358		
		$SDAE_1$	-	-	1	0.0188		
		$SDAE_2$	-	-	-	1		

Table 6.6: Paired ANOVA tests for volume differences between the automatic approaches to segment OARs from group A.

tistically significant across the four groups ($p < 0.05$). Results from volume differences showed more dissimilarities in the paired ANOVA tests than previous metrics (Table 6.6). Paired ANOVA tests pointed out that differences between volumes generated by the SVM_1 framework and volumes generated with the other three settings were statistically significant in nearly all the structures. This situation was almost similarly repeated by the configuration SVM_2, where differences on generated volumes in comparison from those generated by the deep networks were often statistically significant. Last, volumes generated by the two SDAE settings did not present significant differences, with exception of the brainstem, where the paired ANOVA test provided a p-value of 0.0188.

Sensitivity and specificity. Mean sensitivity and specificity values across OARs of group A for the four different classifier configurations are reported in table 6.7. Sensitivity ans specificity obtained with the proposed $SDAE_2$ framework were commonly among the top-ranked results for all the organs of group A. Particularly, sensitivity values achieved by the proposed setting were the highest in the cases of the brainstem and lenses, and among the two highest when segmenting the eyes. Furthermore, standard deviation of sensitivity was reduced on the configurations that employed SDAE as classifier. We can also observe that the inclusion of proposed features into SDAE schemes slightly improved sensitivity values across all the structures. Nevertheless, this trend was not observed in settings employing SVM as classifier. For example, the combination of proposed features with SVM, SVM_2, achieved lower sensitivity values than SVM_1 when segmenting both eyes.

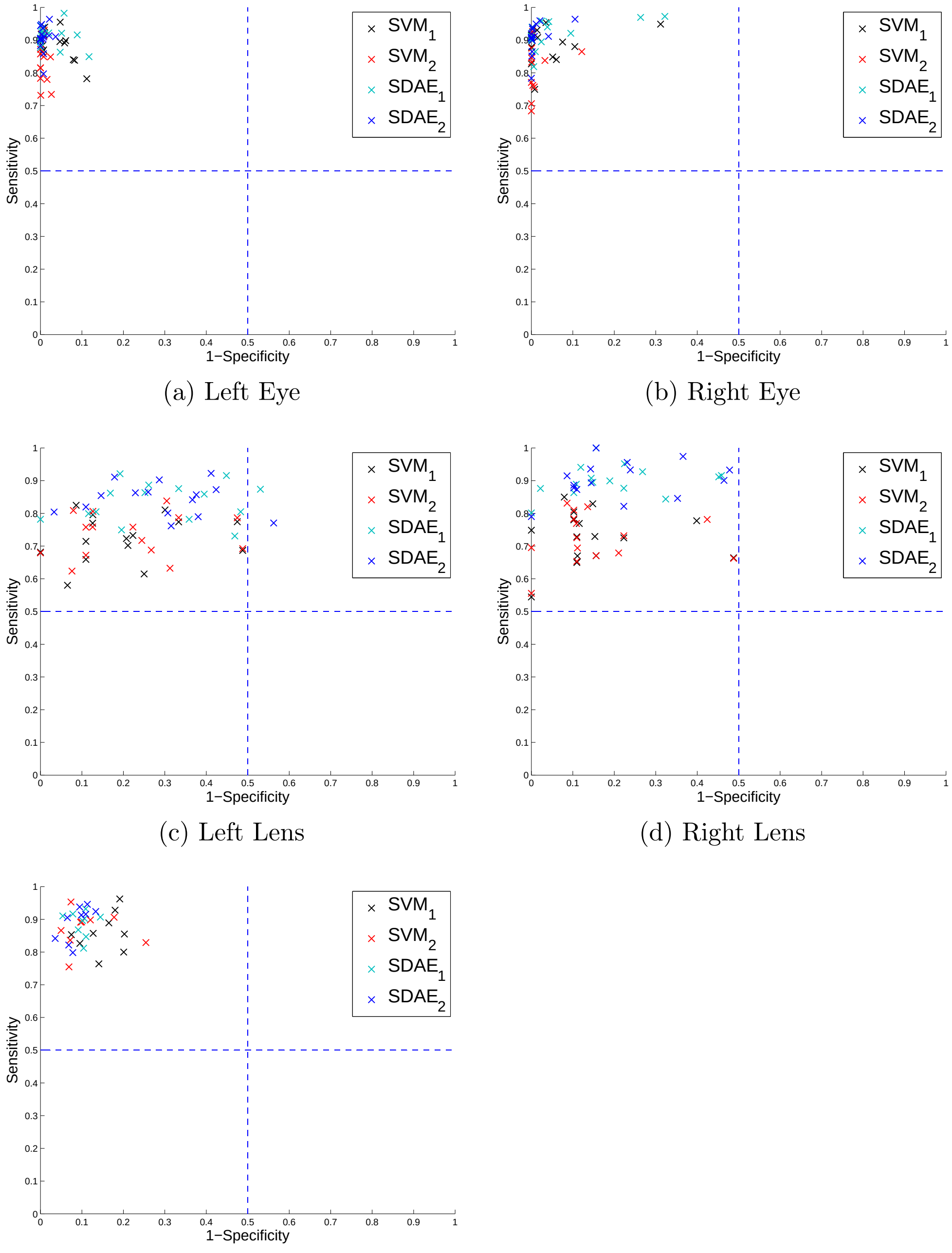
	Configuration	Sensitivity	Specificity
Brainstem	SVM_1	85.94 (± 6.12)	84.69 (± 3.69)
	SVM_2	86.96 (± 5.71)	87.75 (± 6.51)
	SDAE_1	88.67 (± 3.83)	90.01 (± 2.48)
	SDAE_2	90.56 (± 5.41)	91.43 (± 2.97)
Eye (L)	SVM_1	88.63 (± 4.33)	96.49 (± 3.69)
	SVM_2	84.34 (± 6.54)	99.34 (± 0.92)
	SDAE_1	91.25 (± 3.32)	97.29 (± 3.71)
	SDAE_2	91.03 (± 4.02)	99.19 (± 1.06)
Eye (R)	SVM_1	88.79 (± 5.42)	95.46 (± 8.06)
	SVM_2	83.25 (± 7.87)	98.84 (± 3.15)
	SDAE_1	91.74 (± 4.13)	94.47 (± 10.03)
	SDAE_2	90.81 (± 4.62)	98.68 (± 2.78)
Lens (L)	SVM_1	72.29 (± 7.09)	79.26 (± 14.31)
	SVM_2	73.37 (± 6.68)	78.17 (± 14.61)
	SDAE_1	83.40 (± 5.89)	71.08 (± 15.67)
	SDAE_2	84.24 (± 5.01)	70.76 (± 13.77)
Lens (R)	SVM_1	72.94 (± 7.96)	84.68 (± 13.14)
	SVM_2	72.41 (± 7.59)	84.22 (± 13.56)
	SDAE_1	89.98 (± 4.71)	80.39 (± 13.50)
	SDAE_2	90.22 (± 5.65)	79.68 (± 14.31)

Table 6.7: Sensitivity and specificity mean values for the four automatic configurations across the OARs of group A.

In terms of specificity, however, results varied across the four configurations. For large structures, for example, configurations including the proposed features reported the highest specificity values, in comparison with classical features sets. Contrary, for small organs, i.e. lenses, classical features settings achieved marginally higher results than their homologous with proposed features. Differences between SVM and SDAE settings, employing the same features set, mainly come from the brainstem and lenses. Mean specificity values obtained from brainstem segmentations were around 5% higher in SDAE than in SVM configurations. In the case of lenses segmentations, highest specificity values went to the SVM side, which mean values ranging from 5-10% higher than SDAE settings.

A good classifier should ideally be a combination of both high sensitivity and specificity values. We can thereby say that SDAE settings were better classifiers than configurations employing SVM. Additionally, the introduction of proposed features into the classifier improved, although marginally, sensitivity and specificity values of segmentations of OARs from this group.

Following ROC subdivision presented in Section 5.5.1.3, figure 6.9 is presented. On this figure, crosses indicate the correspondence between sensitivity and (1 - specificity) for each patient for the four automatic settings. Thus, each cross represents a single patient and its color indicates the setting employed. It can be observed that for the four analyzed configurations, nearly all results lie on the left-top sub-space, which indicates contours would be considered acceptable for RTP. However, automatic lenses contours for two patients lie on the "high risk" region when employing SDAE_1 and SDAE_2. Furthermore, it is important to note that for the case of both lenses, there are some results that dangerously approach the "high risk" and "poor" regions. While some segmentations generated by SDAE based classifiers are closer to the "high risk" region, segmentations generated by SVM are typically closer to the "poor" region.



(a) Left Eye

(b) Right Eye

(c) Left Lens

(d) Right Lens

(e) Brainstem

Figure 6.9: ROC sub-division analysis for the four automatic approaches for organs of group A.

Some visual examples of automatic segmentations generated by the four settings are shown in Figure 6.10, together with the reference standard. OARs shown in this figure are: eye (left), lens (middle) and brainstem (right). First, it can be observed that in this set of OARs, SVM settings tended to provide larger volumes than those produced by SDAE configurations. Concerning proposed features, their inclusion into the features vector generally produced more similar contours to the reference volume than those generated by schemes incorporating classical features.

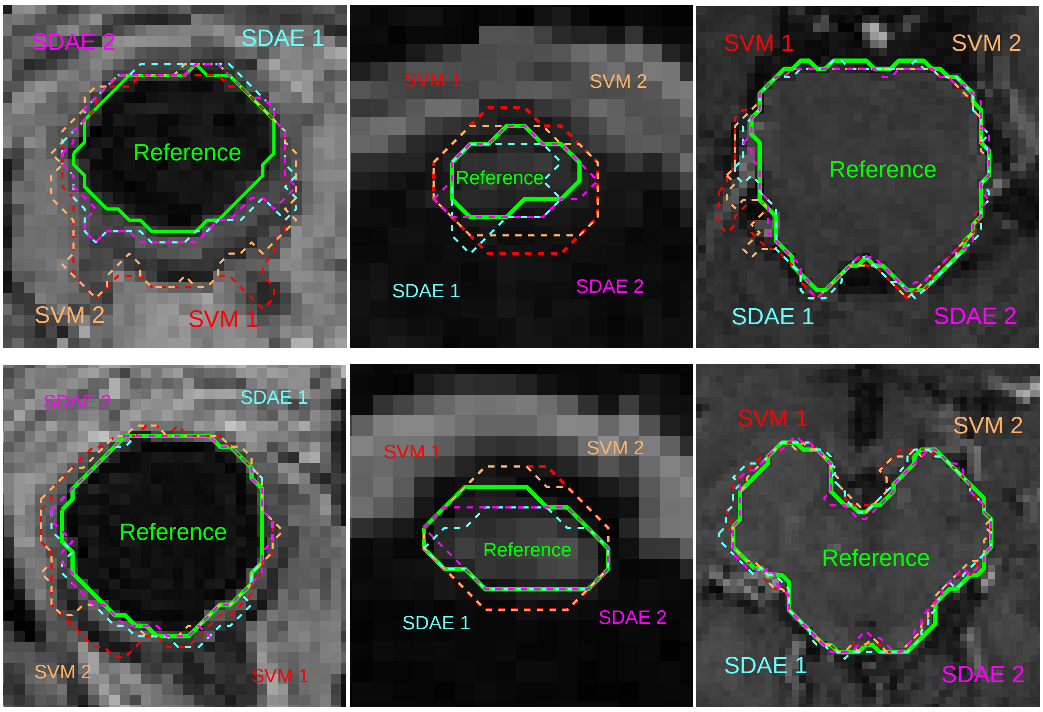


Figure 6.10: Visual examples of automatic segmentations of OARs from group A.

6.2.1.2 Comparison across manual contours and the proposed scheme

This section evaluates manual segmentations in relation with the generated reference standard and compares with the automatic segmentations obtained with our approach. The goal is to quantitatively demonstrate that segmentations generated by our proposed learning scheme lies on the variability of the experts. Since the brainstem was the only structure in group A from which we obtained more than one manual contour per patient, this section only contains results for the brainstem.

Dice Similarity Coefficients. Mean DSC values for the four observers ranged from 0.84 to 0.90, with minimum and maximum values of 0.78 and 0.93 respectively. On the other hand, our proposed approach achieved a mean

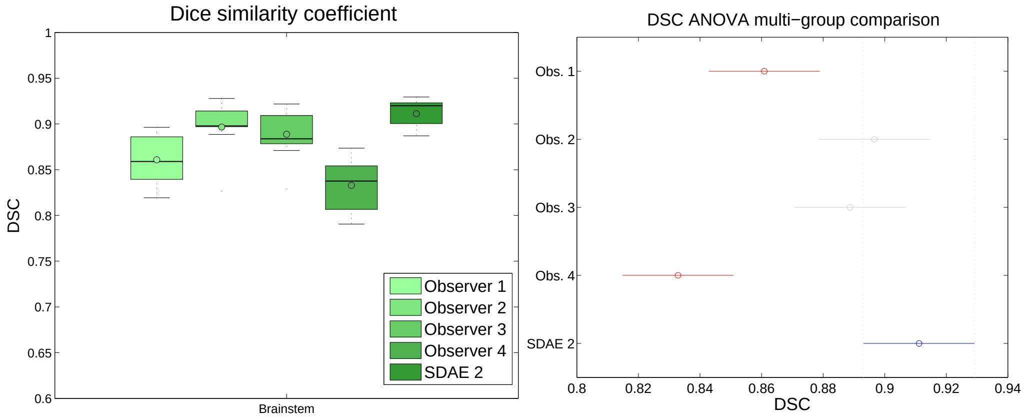


Figure 6.11: DSC results of manual and our proposed approach for the brainstem.

DSC value of 0.92, with a minimum value of 0.89 and a maximum value of 0.93. It can be observed on figure 6.11,left that mean DSC achieved by the proposed system is higher than values reported by manual segmentations when compared with the reference standard. The within-subjects ANOVA test conducted on the DSC of all the groups (p<0.05) indicated that there were significant differences among them. These differences were especially notorious on observers 1 and 4. In the right side of this figure, the ANOVA multi-group comparison is presented. The proposed scheme is represented by a blue line, while groups which have means significantly different from SDAE group are drawn in red. These groups represent to the observer 1 and 4.

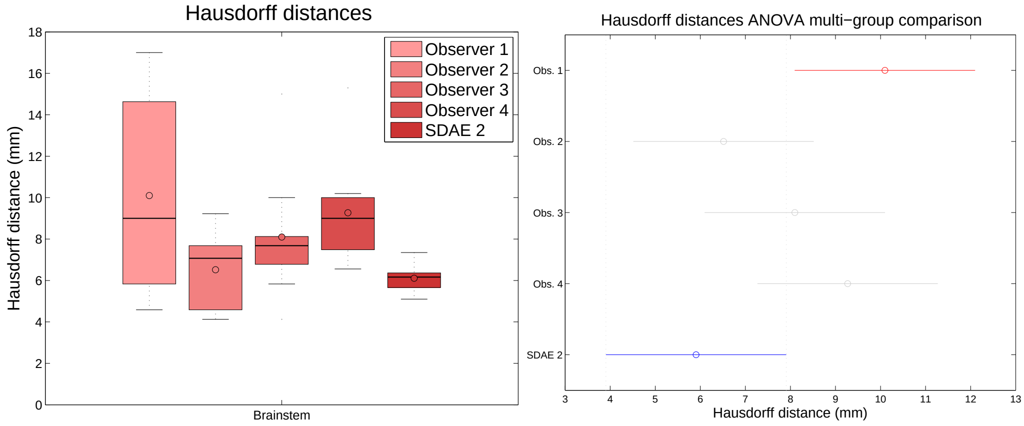


Figure 6.12: Hausdorff distance results of manual and our proposed approach for the brainstem.

Hausdoff distances. Left side of figure 6.12 plots Hausdorff distances distributions for the group of manual and automatic contours. While mean HD values for the four observers ranged from 6.52 to 10.09 mm, our proposed system achieved a mean HD of 5.87 mm. Minimum and maximum HD values obtained by the group of manual raters were 4.12 and 16.93, respectively. Although minimum HD values were not decreased when employing the deep learning scheme, maximum values were reduced to almost the half in relation to several observers. Furthermore, variability of the reported HD was also decreased by the proposed system. The within-subjects ANOVA test conducted on the HD of all the groups indicated that there were not significant differences among them ($p = 0.0225$). However, despite dissimilarities observed across the observers and the automatic approach, only segmentations from observer 1 presented significant differences with respect to automatic contours (Figure 6.12, right). On this figure, groups with means significantly different from our approach, in blue, are displayed in red.

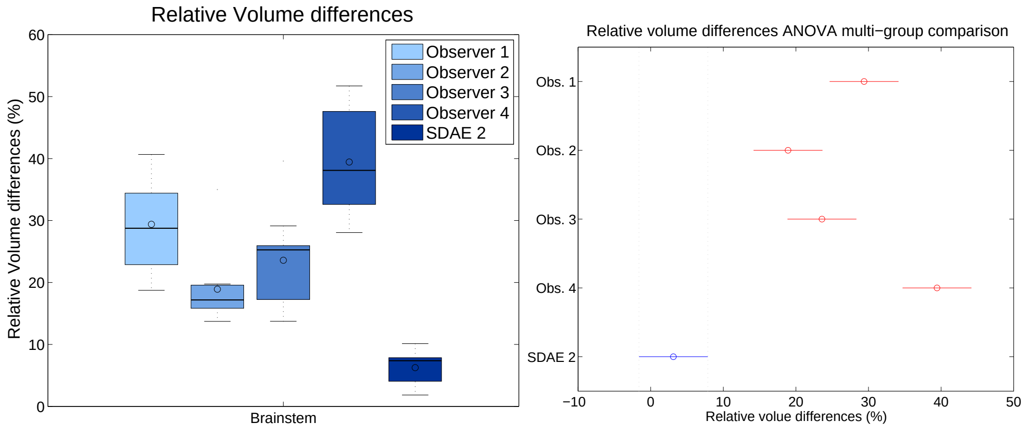


Figure 6.13: Volume differences results of manual and our proposed approach for the brainstem.

Relative volume differences. Mean relative volume differences with regards to reference contours across the four manual observers were reported to be of 29.39%, 18.92%, 23.59% and 39.44%, for observer 1,2,3 and 4, respectively. By employing the deep learning based classification scheme, relative volume difference was reduced to a mean value of 3.10% with respect to reference volume. The within-subjects ANOVA test conducted on volume differences of all the groups indicated that there were significant differences among them ($p = 5.5216x10e^{(-12)}$). These differences come from the manual groups with respect to the automatic method (Figure 6.13). In this figure the ANOVA multi-group comparison for volume differences is shown. The blue

Brainstem			
	DSC	**HD (mm)**	**Abs.Vol.Diff (%)**
Observer 1	0.86 (± 0.03)	10.09 (± 4.84)	29.39 (± 8.26)
Observer 2	0.90 (± 0.03)	6.52 (± 1.99)	18.92 (± 6.45)
Observer 3	0.88 (± 0.03)	8.09 (± 3.06)	23.59 (± 7.99)
Observer 4	0.84 (± 0.03)	9.27 (± 2.64)	39.44 (± 8.54)
Our method	0.92 (± 0.02)	5.87 (± 0.73)	3.10 (± 1.18)

Table 6.8: Comparisons across the four observers and the proposed approach when segmenting the brainstem.

line represents the automatic SDAE setting. Red lines symbolize the group comprising the manual raters. As it can be observed, mean of SDAE have significant differences with respect to all the raters of manual segmentations group.

Table 6.8 summarizes the performance of manual annotations of the brainstem done by the four observers in comparison with the proposed approach. For the three metrics, the proposed approach significantly outperforms manual annotations, particularly in terms of relative volume differences.

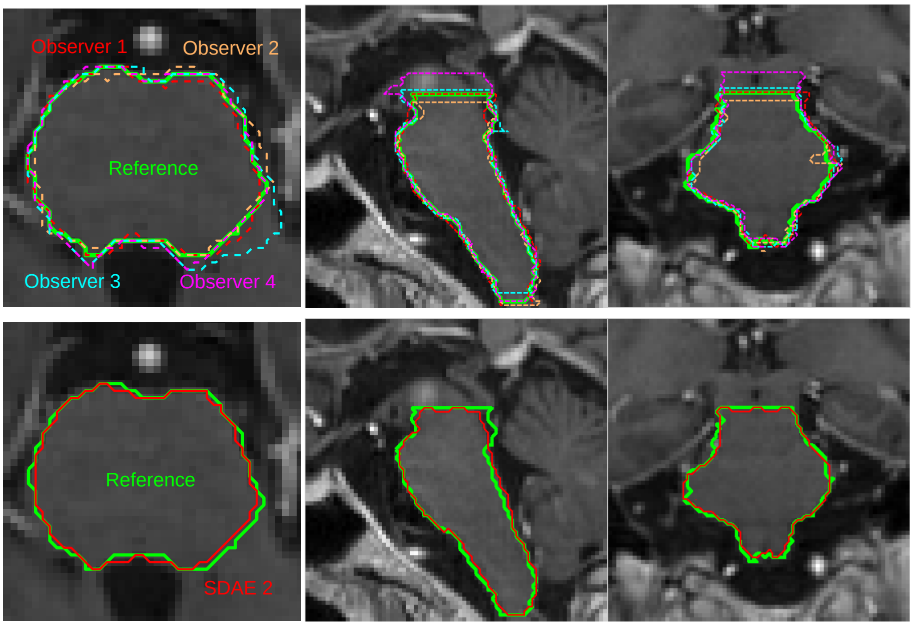


Figure 6.14: Visual examples of manual brainstem delineation and their comparison with reference and automatic segmentations.

Figure 6.14 shows a visual example of manual contours (*top*) and contours generated by our approach (*bottom*) when segmenting the brainstem, and its comparison with the reference standard. It can be observed from the manual contours that differences between manual raters usually come from the z axis and from areas where no visible anatomical boundaries exist.

6.2.2 OARs group B

This section presents results of the automatic approaches to segment OARs that belong to group B: optic nerves, pituitary gland, pituitary stalk and chiasm. As in section 6.2.1, organs separately present in both left and right brain sides are split into the two sections, which are individually analyzed. With regards to the features sets employed, we refer to the table 6.2, where different groups of features were presented. Following this definition, the deep learning scheme that employs classical features will be referred to as SDAE$_1$. The rest of the groups will be referred to as SDAE$_{Augmented}$, SDAE$_{Textural}$ and SDAE$_{AE-FV}$, for the augmented, textural and AE-FV set, respectively. As in the previous section, and to investigate the impact of employing a deep network as classifier instead of some other classification schemes, SVM is used as reference. Both the classical and the AE-FV configurations in combination with SVM will be included in the evaluation. These settings will be referred to as SVM$_1$ and SVM$_{AE-FV}$, respectively.

6.2.2.1 Comparison with respect to the reference standard

Dice Similarity Coefficients. Dice similarity coefficients obtained with the automatic segmentations with respect to the reference standard for the OARs of group B are plotted in Figure 6.15. Box plots are grouped for each OAR. Inside each group, results for SVM references, and the several SDAE settings are displayed. Among all configurations, SVM based classifiers presented the lowest overall mean DSC values, with 0.59 (± 0.16) and 0.64 (± 0.09) for SVM$_1$ and SVM$_{AE-FV}$, respectively. Concerning the SDAE settings, the system that included our proposed features, SDAE$_{AE-FV}$, achieved the highest mean DSC value over all the OARs. Values for the several SDAE configurations were: 0.69 (± 0.11), 0.74 (± 0.07), 0.74 (± 0.07) and 0.79 (± 0.06), for classical, augmented, textural and AE-FV sets, respectively. Analyzing each structure separately, we can observe that again, mean DSC values from SVM configurations were among the lowest ones. In this setting, adding the set of proposed features generally improved the mean DSC. Nevertheless, it often remained below mean values achieved by SDAE based classifiers.

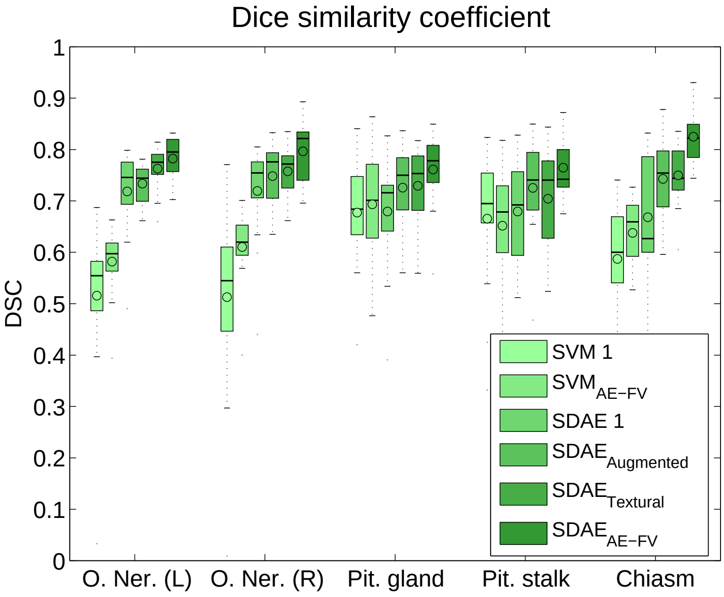


Figure 6.15: Segmentation DSC results for the automatic contours with different settings for organs of group B.

Regarding the impact of different features sets on deep architectures, the use of classical features produced segmentations with acceptable mean DSC across all the OARs. However, it did not improve any of the other three features groups. Mean DSC for SDAE$_1$ were 0.72 (± 0.09), 0.72 (± 0.10), 0.68 (± 0.12), 0.68 (± 0.10) and 0.67 (± 0.13) for left optic nerve, right optic nerve, pituitary gland, pituitary stalk and chiasm, respectively. Introduction of either augmented or textural features improved the segmentation performance of the classifier, which is reflected on its mean DSC values. In the same order, mean DSC values were 0.73 (± 0.04), 0.75 (± 0.06), 0.73 (± 0.08), 0.73 (± 0.09) and 0.74 (± 0.08) for the augmented features set, and 0.76 (± 0.05), 0.76 (± 0.06), 0.73 (± 0.08), 0.70 (± 0.10) and 0.75 (± 0.06) when employing the textural features set. Last, the use of the proposed features set, i.e. AE-FV, achieved the highest mean DSC values across all the structures with values of 0.78 (± 0.05), 0.80 (± 0.06), 0.76 (± 0.06), 0.77 (± 0.08) and 0.83 (± 0.06), respectively.

Automatic segmentations presented significant differences ($p < 0.05$) across the automatic groups, according to the within-subjects ANOVA test on the DSC of all the groups. Paired repeated measures ANOVAs were conducted over groups that employed only classical and proposed features. The objective of performing paired repeated ANOVAs only in classical and proposed features was to evaluate whether the inclusion of proposed features set in this thesis made a significant difference with respect to classical features set. Results of

Paired ANOVA (DSC)					
		SVM$_1$	**SVM$_{AE-FV}$**	**SDAE$_1$**	**SDAE$_{AE-FV}$**
Optiv Nerve (L)	**SVM$_1$**	1	0.1386	0.0001	5.9524x10^{-7}
	SVM$_{AE-FV}$	-	1	3.8499x10^{-5}	2.1743x10^{-6}
	SDAE$_1$	-	-	1	0.0159
	SDAE$_{AE-FV}$	-	-	-	1
Optiv Nerve (R)	**SVM$_1$**	1	0.0737	0.0008	2.8712x10^{-7}
	SVM$_{AE-FV}$	-	1	0.0015	7.8942x10^{-6}
	SDAE$_1$	-	-	1	0.0138
	SDAE$_{AE-FV}$	-	-	-	1
Pituitary Gland	**SVM$_1$**	1	0.6793	0.9564	0.0173
	SVM$_{AE-FV}$	-	1	0.7341	0.0472
	SDAE$_1$	-	-	1	0.0291
	SDAE$_{AE-FV}$	-	-	-	1
Pituitary Stalk	**SVM$_1$**	1	0.7635	0.7507	0.0147
	SVM$_{AE-FV}$	-	1	0.4761	0.0014
	SDAE$_1$	-	-	1	0.0081
	SDAE$_{AE-FV}$	-	-	-	1
Chiasm	**SVM$_1$**	1	0.1503	0.0807	9.1865x10^{-9}
	SVM$_{AE-FV}$	-	1	0.4281	8.4951x10^{-8}
	SDAE$_1$	-	-	1	0.0002
	SDAE$_{AE-FV}$	-	-	-	1

Table 6.9: Paired ANOVA tests for the DSC between the automatic approaches to segment OARs from group B.

these tests on DSC values are presented in table 6.9, which shows p-values obtained when comparing results between only two groups. Results demonstrate that no statistically significant differences existed between both SVM based systems in any of the OARs of this group ($p > 0.05$). Regarding the use of deep networks, the combination of SDAE as classifier with classical features reported significant differences with respect to SVM groups when segmenting both optic nerves. Our proposed scheme, however, presented differences on DSC values that were statistically significant with respect the other groups in all the OARs.

Hausdorff distances. Figure 6.16 plots the distribution of HD across the OARs for all the automatic frameworks. As in the case of DSC distributions, mean HD values over all the structures show that SVM based classifiers presented the worst results. While SVM$_1$ and SVM$_{AE-FV}$ achieved an overall mean HD of 7.09 (± 5.23) and 6.63 (± 5.09) mm, respectively, mean values for SDAE settings were 5.80 (± 5.47), 4.74 (± 4.83), 4.69 (± 4.70) and 3.32 (± 0.96) mm for SDAE$_1$, SDAE$_{Augmented}$, SDAE$_{Textural}$ and SDAE$_{AE-FV}$, respectively. Looking at each structure individually, it can be observed that including the set of proposed features into the SVM system decreased mean HD values with respect to the classical features set when segmenting both optic nerves. For the rest of the organs, however, inclusion of proposed features did not particularly improve HD values. Mean HD values for left optic nerve, right optic nerve, pituitary gland, pituitary stalk and chiasm, were reported to

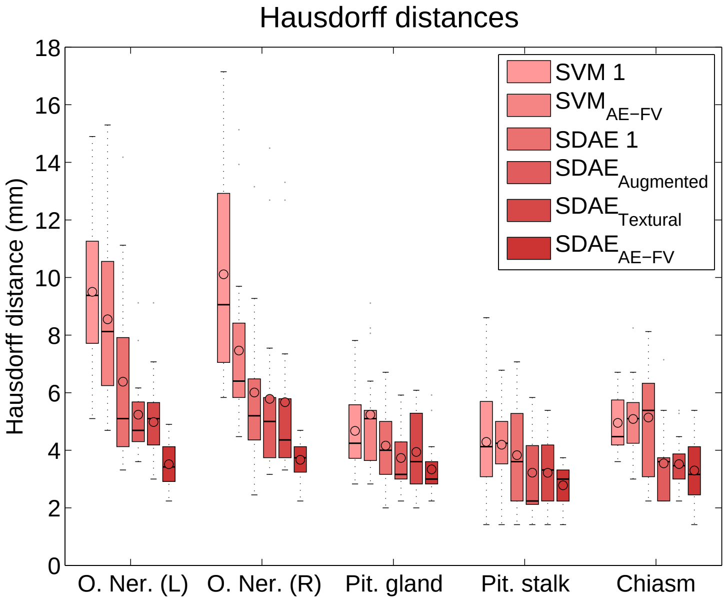


Figure 6.16: Segmentation HD results for the automatic contours with different settings for organs of group B.

be of 11.47 (± 8.12), 10.11 (± 3.89), 4.67 (± 1.45), 4.29 (± 1.97) and 4.95 (± 1.02) mm for SVM$_1$, and 11.18 (± 9.22), 7.46 (± 3.23), 5.24 (± 1.99), 4.20 (± 1.38) and 5.86 (± 1.31) mm, respectively, for SVM$_{AE-FV}$. Employing SDAE as classifier instead of SVM in a classical features setting decreased mean HD in most cases. Incorporation of either augmented or textural features in the SDAE based classifier improved HD values with respect to classical features. While in some organs mean HD values were lower for augmented features based classifiers, for some other organs textural features set achieved the lowest mean HD values. Nevertheless, the combination of both features sets into the AE-FV set led to the lowest mean HD values across all the structures. Mean HD values obtained with the proposed features set were 3.51 (± 0.87), 3.67 (± 0.67), 3.34 (± 1.09), 2.78 (± 0.76) and 3.29 (± 1.19), for left and right optic nerve, pituitary gland, pituitary stalk and chiasm, respectively.

Paired repeated measures ANOVAs conducted on HD values (Table 6.10) indicates that including proposed features in the SVM based classifier did not produce segmentations with significant differences with respect to classical configurations ($p > 0.05$). Employing SDAE as classifier with the classical features set did not report differences statistically significant in four out five structures. Only segmentations of the right optic nerve ($p = 0.0098$) showed significant different between SVM and SDAE when employing the classical features set. Nevertheless, differences between the two classifiers, i.e. SVM and SDAE, were significant when employing proposed features over all the structures ($p < 0.05$). Regarding the use of proposed features against classical

Paired ANOVA (Hausdorff distances)					
		SVM$_1$	**SVM$_{AE-FV}$**	**SDAE$_1$**	**SDAE$_{AE-FV}$**
Optiv Nerve (L)	**SVM$_1$**	1	0.9318	0.5786	0.0014
	SVM$_{AE-FV}$	-	1	0.4435	0.0034
	SDAE$_1$	-	-	1	0.0377
	SDAE$_{AE-FV}$	-	-	-	1
Optiv Nerve (R)	**SVM$_1$**	1	0.0519	0.0098	7.7642x10^{-7}
	SVM$_{AE-FV}$	-	1	0.3869	0.0001
	SDAE$_1$	-	-	1	0.0057
	SDAE$_{AE-FV}$	-	-	-	1
Pituitary Gland	**SVM$_1$**	1	0.3855	0.3358	0.0077
	SVM$_{AE-FV}$	-	1	0.1008	0.0031
	SDAE$_1$	-	-	1	0.0836
	SDAE$_{AE-FV}$	-	-	-	1
Pituitary Stalk	**SVM$_1$**	1	0.8769	0.5265	0.0099
	SVM$_{AE-FV}$	-	1	0.5917	0.0017
	SDAE$_1$	-	-	1	0.0616
	SDAE$_{AE-FV}$	-	-	-	1
Chiasm	**SVM$_1$**	1	0.7512	0.7921	0.0003
	SVM$_{AE-FV}$	-	1	0.9461	0.0005
	SDAE$_1$	-	-	1	0.0165
	SDAE$_{AE-FV}$	-	-	-	1

Table 6.10: Paired ANOVA tests for the Hausdorff distances between the automatic approaches to segment OARs from group B.

features in SDAE settings, segmentation of both optic nerves and chiasm presented significant differences between them, with p-values of 0.0377, 0.0057 and 0.0165, respectively.

Relative Volume Differences. Distributions of relative volume differences of the six automatic schemes for each organ are plotted in figure 6.17. Schemes employing SVM as classifier presented the largest volume differences for all the OARs of group B. Indeed, with exception of the pituitary stalk, mean relative volume differences for SVM based system were double than those reported by SDAE settings, independently on the features set used. Taking results from each structure, it can be observed that by employing either augmented or textural features in SDAE settings did not reduce mean rVD with respect to classical features. Actually, in some cases, such as both optic nerves, differences in volume were higher when employing one of these groups. However, the proposed features set, which comprises all these groups, achieved the lowest rVD among all the configurations. Mean values for relative volume differences across the six groups for the 6 OARs follows. The order of the OARs is: left optic nerve, right optic nerve, pituitary gland, pituitary stalk and chiasm. For SVM$_1$ mean rVD were 72.58 (± 22.86), 72.14 (± 42.59), 53.37 (± 48.89), 23.44 (± 15.16) and 83.24 % (± 84.49). For SVM including the proposed AE-FV set: 52.86(± 15.46), 41.48 (± 14.69), 71.49 (± 51.68), 38.10 (± 32.55) and 79.28 (± 43.64). First of SDAE configurations, which employed classical features, obtained the following mean values: 22.06 (± 13.92), 15.10

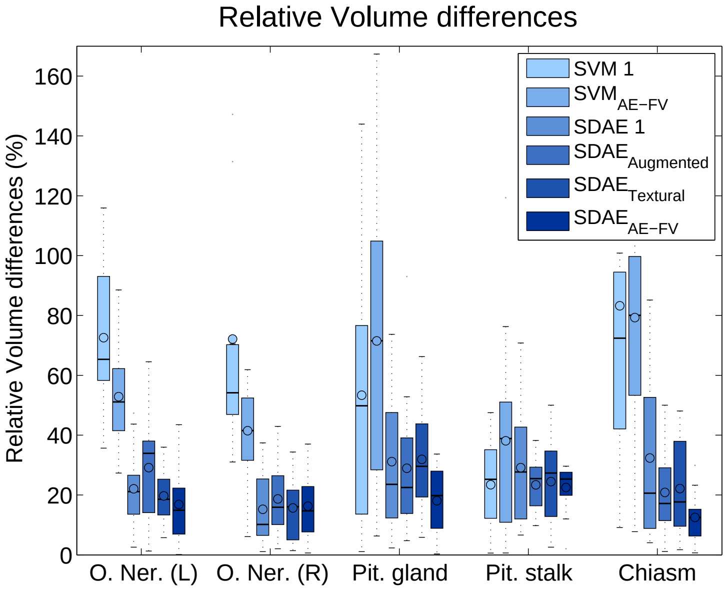


Figure 6.17: Relative volume differences results for the automatic contours with different settings for organs of group B.

(± 11.62), 31.14 (± 23.82), 29.13 (± 20.41) and 32.37 (± 27.58). When incorporating augmented features into the features set, mean rVD were: 29.13 (± 18.08), 18.67 (± 12.42), 28.95 (± 23.42), 23.38 (± 8.82) and 20.85 (± 14.77). If we employed textural features instead, mean values of rVD were: 19.68 (± 9.56), 15.68 (± 11.06), 31.89 (± 18.29), 24.46 (± 14.26) and 22.14 (± 15.34). Finally, our proposed system achieved the following mean rVD values: 16.85 (± 13.39), 16.27 (± 11.09), 18.09 (± 11.29), 22.51 (± 7.55) and 12.48 (± 7.69).

Automatic segmentations presented significant ($p < 0.05$) differences across the automatic groups. Paired repeated measures ANOVAs (Table 6.11) indicate that differences between groups, in terms of volume differences, were significant in most of the cases. Results from the SVM based scheme that employed proposed features were significantly different from those obtained by the classical setting when segmenting both optic nerves and pituitary stalk. Concerning the use of SDAE as classifier, results from SDAE settings were significant different than SVM settings in all the organs, with exception of the pituitary stalk. In this case, with p-values of 0.0961 and 0.7652 for SDAE_1 and SDAE_{AE-FV}, respectively, differences on volume were not statistically significant between SVM_1 and both SDAE groups. On the other hand, the impact of adding proposed features into the deep learning scheme was statistically significant only when segmenting the pituitary stalk and chiasm (p=0.0394 and p=0.0068), in terms of volume differences.

Paired ANOVA (Relative volume differences)					
		SVM$_1$	**SVM$_{AE-FV}$**	**SDAE$_1$**	**SDAE$_{AE-FV}$**
Optiv Nerve (L)	**SVM$_1$**	1	0.0099	3.2411x10^{-8}	9.3889x10^{-9}
	SVM$_{AE-FV}$	-	1	1.3548x10^{-6}	3.2796x10^{-7}
	SDAE$_1$	-	-	1	0.2633
	SDAE$_{AE-FV}$	-	-	-	1
Optiv Nerve (R)	**SVM$_1$**	1	0.0135	8.4021x10^{-6}	1.0771x10^{-5}
	SVM$_{AE-FV}$	-	1	4.1064x10^{-6}	7.0974x10^{-6}
	SDAE$_1$	-	-	1	0.8851
	SDAE$_{AE-FV}$	-	-	-	1
Pituitary Gland	**SVM$_1$**	1	0.3047	0.0114	0.0024
	SVM$_{AE-FV}$	-	1	0.0006	6.7363x10^{-5}
	SDAE$_1$	-	-	1	0.7741
	SDAE$_{AE-FV}$	-	-	-	1
Pituitary Stalk	**SVM$_1$**	1	0.0004	0.0961	0.7652
	SVM$_{AE-FV}$	-	1	6.5727x10^{-6}	0.0005
	SDAE$_1$	-	-	1	0.0394
	SDAE$_{AE-FV}$	-	-	-	1
Chiasm	**SVM$_1$**	1	0.9514	0.0002	0.0021
	SVM$_{AE-FV}$	-	1	2.4641x10^{-8}	2.4966x10^{-7}
	SDAE$_1$	-	-	1	0.0068
	SDAE$_{AE-FV}$	-	-	-	1

Table 6.11: Paired ANOVA tests for volume differences between the automatic approaches to segment OARs from group B.

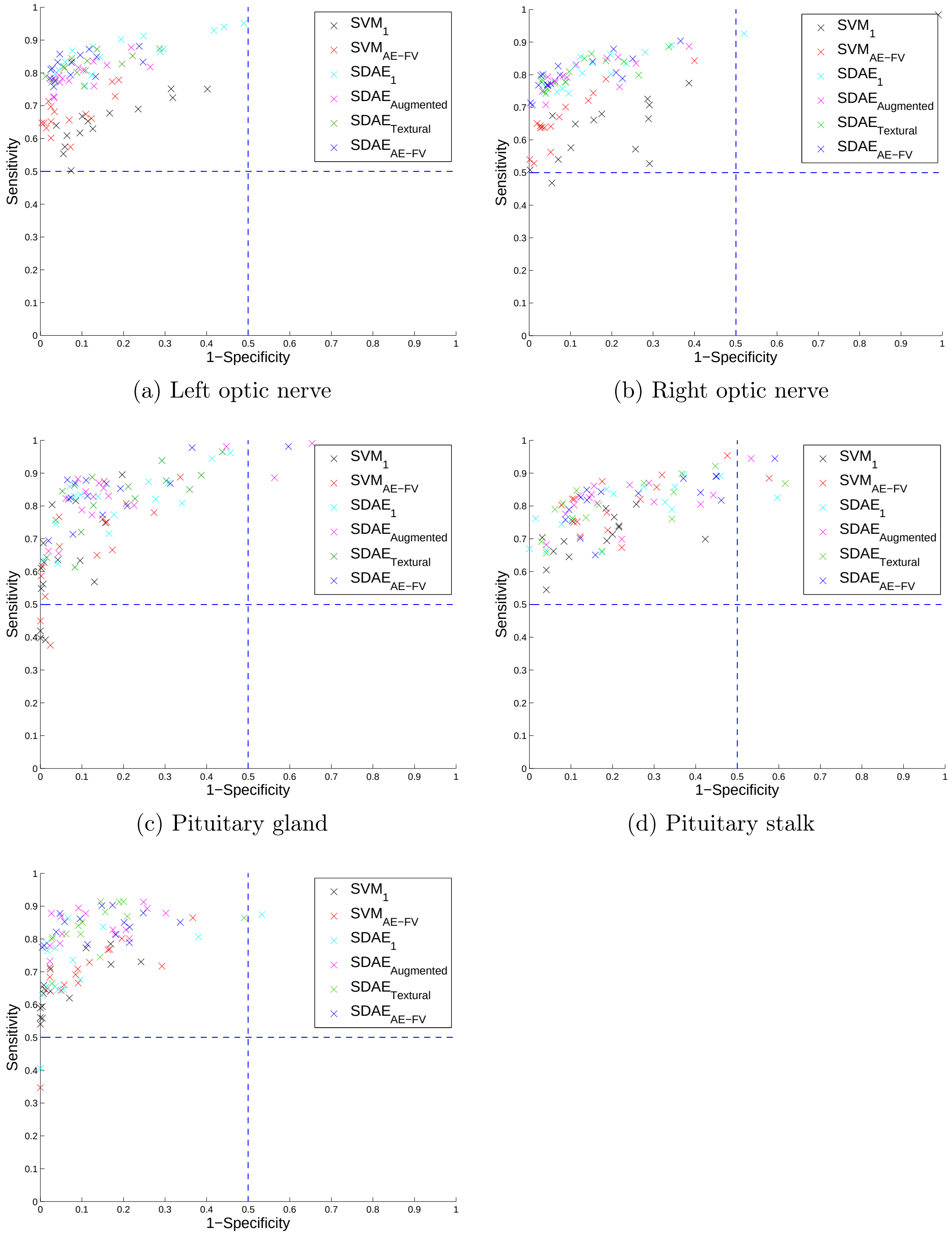
Sensitivity and specificity. Sensitivity and specificity across OARs of group B for the six different classifier configurations are reported in table 6.12. In general, SDAE based classifiers achieved the highest sensitivity values, whereas SVM settings obtained the highest specificity rates. Mean sensitivity values for both SVM configurations commonly ranged between 60 and 70, with exception of the pituitary stalk, where sensitivity was around 70 for SVM$_1$ and close to 80 for SVM$_{AE-FV}$. Employing the SDAE system with classical features improved sensitivity, leading to values close to 80 for all the organs with exception of the chiasm, which mean sensitivity value was 71.67. Adding any single of the investigated features set (SDAE$_{Augmented}$ or SDAE$_{Textural}$) typically increased sensitivity with respect to classical settings, with mean values nearly 80, or little bit higher. At last, the proposed system achieved sensitivity values greater than 80 in all the structures. Contrary, concerning the specificity, any pattern was noticed. For instance, regarding the use of SVM with classical or proposed features, specificity was increased when segmenting both optic nerves, whilst it was decreased when segmenting pituitary stalk or chiasm.

Combination of higher sensitivity and specificity metrics obtained from the AE-FV based classifier indicated that the proposed system correctly identified more tissue voxels than the others settings did, and also was better at rejecting tissue voxels that were not related to the tissue class of interest.

The performance of the automatic delineations according the features set

	Configuration	**Sensitivity**	**Specificity**
Optic nerve (L)	SVM$_1$	66.68 (± 10.74)	79.19 (± 23.57)
	SVM$_{AE-FV}$	67.46 (± 5.69)	92.86 (± 6.64)
	SDAE$_1$	85.41 (± 5.76)	79.38 (± 15.07)
	SDAE$_{Augmented}$	79.18 (± 4.01)	90.44 (± 7.27)
	SDAE$_{Textural}$	81.87 (± 3.49)	89.34 (± 7.65)
	SDAE$_{AE-FV}$	82.23 (± 3.71)	91.02 (± 7.31)
Optic nerve (R)	SVM$_1$	64.74 (± 12.81)	76.52 (± 23.82)
	SVM$_{AE-FV}$	66.31 (± 8.68)	91.29 (± 10.32)
	SDAE$_1$	79.30 (± 6.13)	82.79 (± 13.28)
	SDAE$_{Augmented}$	80.53 (± 5.18)	87.67 (± 10.82)
	SDAE$_{Textural}$	80.19 (± 4.84)	87.86 (± 10.86)
	SDAE$_{AE-FV}$	81.54 (± 4.45)	88.09 (± 9.52)
Pituitary gland	SVM$_1$	62.31 (± 15.18)	94.84 (± 6.52)
	SVM$_{AE-FV}$	67.81 (± 14.89)	88.51 (± 10.62)
	SDAE$_1$	80.85 (± 9.69)	80.86 (± 14.32)
	SDAE$_{Augmented}$	83.13 (± 9.29)	79.85 (± 19.35)
	SDAE$_{Textural}$	82.24 (± 10.05)	81.07 (± 13.79)
	SDAE$_{AE-FV}$	84.22 (± 7.94)	82.69 (± 15.09)
Pituitary stalk	SVM$_1$	70.33 (± 6.94)	84.42 (± 10.62)
	SVM$_{AE-FV}$	80.78 (± 7.76)	77.61 (± 14.54)
	SDAE$_1$	79.19 (± 8.02)	76.52 (± 17.42)
	SDAE$_{Augmented}$	81.66 (± 6.47)	77.29 (± 14.28)
	SDAE$_{Textural}$	79.62 (± 8.17)	77.98 (± 17.19)
	SDAE$_{AE-FV}$	82.28 (± 7.53)	73.14 (± 16.86)
Chiasm	SVM$_1$	65.09 (± 7.78)	94.37 (± 7.88)
	SVM$_{AE-FV}$	69.74 (± 11.39)	88.43 (± 10.57)
	SDAE$_1$	71.67 (± 12.07)	89.84 (± 15.23)
	SDAE$_{Augmented}$	83.93 (± 5.16)	86.64 (± 9.69)
	SDAE$_{Textural}$	84.32 (± 7.40)	82.42 (± 17.78)
	SDAE$_{AE-FV}$	83.94 (± 4.34)	86.11 (± 9.71)

Table 6.12: Sensitivity and specificity mean values for the six automatic configurations across the OARs of group B.



(a) Left optic nerve

(b) Right optic nerve

(c) Pituitary gland

(d) Pituitary stalk

(e) Optic chiasm

Figure 6.18: ROC sub-division analysis for the six automatic approaches for organs of group A.

employed is also compared by using ROC region analysis (Fig. 6.18). On this figure, crosses indicate the correspondence between sensitivity and (1 - specificity) for each patient for the six automatic settings. Therefore, each cross represents a single patient and its color indicates the setting employed. First, it can be observed that for the six configurations nearly all results lie on the left-top sub-space, which indicates contours would be considered acceptable for RTP. Nevertheless, there are cases which should be taken into consideration. Some contours generated by SDAE$_1$ approach, or are inside, the "high risk" area when segmenting both optic nerves. In addition, although contours provided by both SVM configurations lie inside the "acceptable" area, they dangerously surround the "poor" region, where the OARs are not spared. Automatic segmentations of pituitary gland and pituitary stalk, from all the settings, also presented some contours that lie outside the acceptable region. Two pituitary gland contours from SDAE$_{Augmented}$, one from SDAE$_{AE-FV}$ and two from both SVM$_1$ and SVM$_{AE-FV}$ were in the "poor" and "high risk" regions. Again, several contours generated by both SVM settings were very close to the "poor" region. In the case of the pituitary stalk, one contour for each SDAE configuration and one from SVM$_{AE-FV}$ were found in the "high risk" region. However, more automatic contours from several settings were close to the line that divided the "acceptable" and "high risk" region. Last, only few contours generated by both settings employing classical features, SVM$_1$ and SDAE$_1$, were not in the "acceptable" region when segmenting the chiasm. As in some previous cases, automatic contours generated by SVM$_1$ were very close to the "poor" region.

Figure 6.19 displays the automatic contours generated by the evaluated configurations. To investigate the effect on the segmentation of employing different classifiers, segmentations from configurations employing either SVM or SDAE are presented on the top-row of this figure. Visual results show that SVM based classifiers provided contours much larger than the reference. This was particularly noticeable in the contours from SVM setting employing classical features. In the case of the chiasm, for example, SVM configurations were not capable of distinguish between chiasm and pituitary stalk. Contrary, classifiers based on SDAE correctly classify the chiasm avoiding the neighboring region of the pituitary stalk. Comparison of the impact on the segmentation performance when adding the different features sets on the SDAE settings can be seen in the bottom-row. Including either augmented or textural features into the classification system typically improved segmentations respect to classical features. Nevertheless, combining all features into the AE-FV set achieved the best contours among the SDAE frameworks.

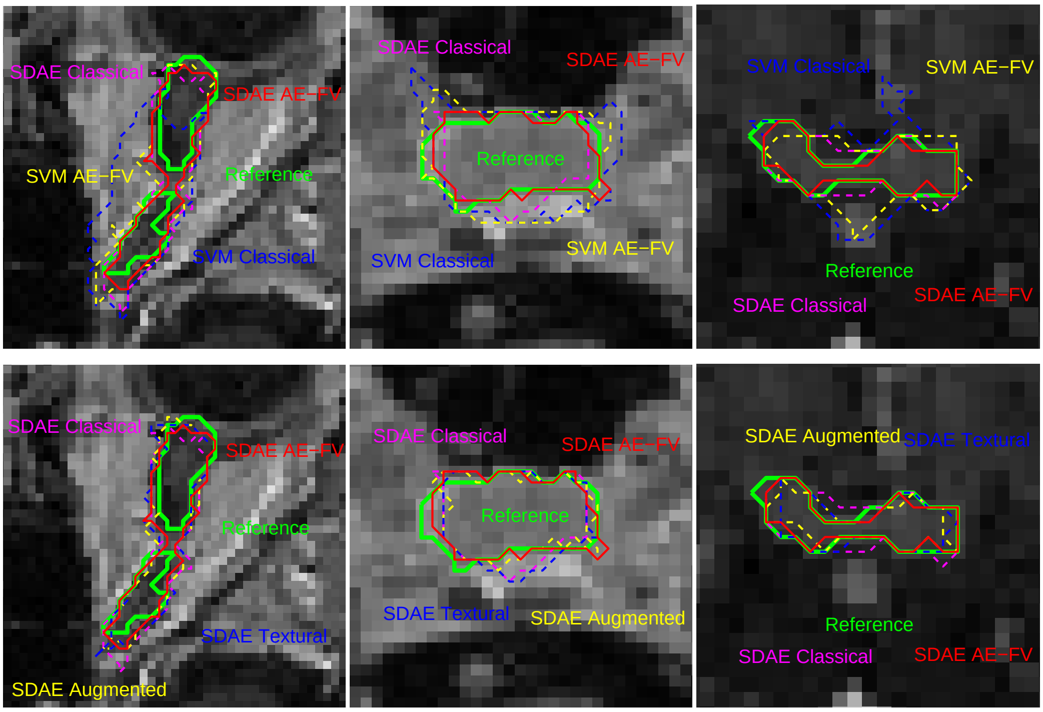


Figure 6.19: Segmentation results produced by the proposed classification system when segmenting the right optic nerve (*left*), pituitary gland (*middle*) and chiasm (*right*), and comparison with the other automatic configurations.

6.2.2.2 Comparison across manual contours and the proposed scheme

As for the case of OARs of group A, this section evaluates manual segmentations in relation with the reference standard and compares with the automatic segmentation obtained with our approach. The goal is again to quantitatively demonstrate that segmentations generated by our proposed learning scheme lies on the variability of the experts. Additionally, in cases where performance of automatic segmentation does not lie on the expert variability, we aim at demonstrating that no significant differences exist between the manual raters and the contours generated by our approach.

Dice Similarity Coefficients. Dice similarity coefficients distribution for the three observers and our proposed approach across all the OARs of group B are plotted in Figure 6.20. Each box group contains several columns representing distributions of the segmentations results for a given organ. While the three first columns of each box group represent to the manual raters, the last columns represent our automatic method. Mean DSC over all the OARs of group B is distributed as follows: 0.83(± 0.07) for observer 1, 0.75(± 0.09)

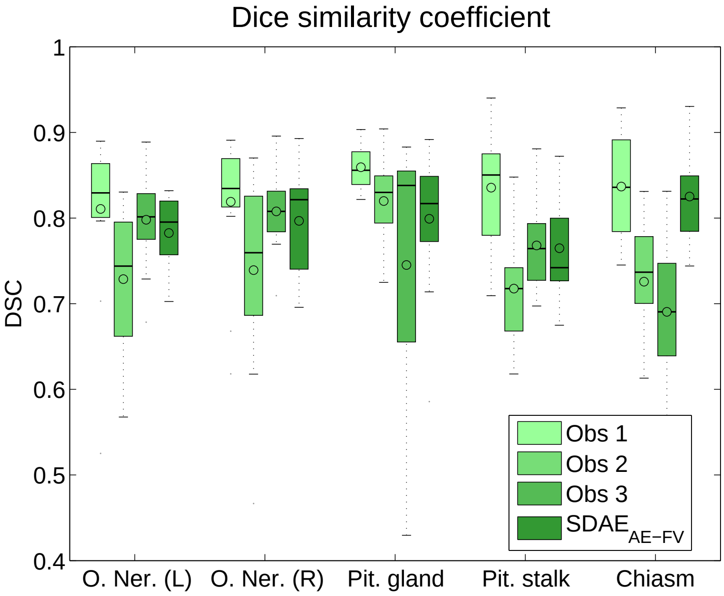


Figure 6.20: DSC results of manual and our proposed approach for the OARs of group B.

ANOVA analysis (DSC)					
	Within-subjects ANOVA	**Paired ANOVA**			
	All		Obs 1	Obs 2	Obs 3
Optic nerve L	0.0108	SDAE$_{AE-FV}$	0.2946	0.0281	0.3881
Optic nerve R	0.0259	SDAE$_{AE-FV}$	0.1158	0.0243	0.1327
Pituitary gland	0.0081	SDAE$_{AE-FV}$	0.0083	0.3913	0.2247
Pituitary stalk	0.0001	SDAE$_{AE-FV}$	0.0035	0.0385	0.8671
Chiasm	0.0001	SDAE$_{AE-FV}$	0.6047	0.0018	0.0001

Table 6.13: P-values of the ANOVA for Dice similarity coefficient results of the OARs of group B.

for observer 2, 0.76(± 0.10) for observer 3 and 0.79(± 0.06) for our automatic approach. Looking at individual structures, mean DSC values obtained from segmentations made by observer 1 were 0.81(± 0.09), 0.82(± 0.08), 0.86(± 0.03), 0.84(± 0.06) and 0.84(± 0.06), for left optic nerve, right optic nerve, pituitary gland, pituitary stalk and chiasm, respectively. In the same order, mean DSC values were 0.73(± 0.08), 0.74(± 0.11), 0.82(± 0.05), 0.72(± 0.06) and 0.73(± 0.09) for segmentations of observer 2 and 0.80(± 0.05), 0.81(± 0.04), 0.75(± 0.15), 0.77(± 0.05) and 0.69(± 0.10) for segmentations of observer 3. Last, our proposed system achieved mean DSC values of 0.78(± 0.05), 0.80(± 0.06), 0.80(± 0.08), 0.77(± 0.08) and 0.83(± 0.06), respectively. It can be observed on figure 6.20 that mean DSC achieved by the proposed system is always between the highest and lowest values reported by manual segmentations when compared with the reference standard.

Results from the ANOVA analysis conducted between all groups together, as well as between results provided by the proposed automatic approach and each of the manual raters are presented in Table 6.13. The within-subjects ANOVA tests conducted on the DSC values of all the groups indicated that there were significant differences among them ($p < 0.05$) in all the OARs. Values from the paired ANOVA tests indicated that there were not significant differences on DSC values between observer 3 and our method in four out of five OARs. Only DSC results from the chiasm presented significant differences between observer 3 and our method. However, if we look at the mean DSC distributions (Figure 6.20), we can observe that in this case our approach outperformed the performance of observer 3, in terms of DSC. Significant differences ($p < 0.05$) between observer 2 and our approach come from the segmentation of left and right optic nerves, pituitary stalk and chiasm. Nevertheless, and as in the previous case, mean DSC distributions (Figure 6.20) indicated that in these cases our approach outperformed the performance of observer 2, in terms of DSC. Regarding the comparison with observer 1, although DSC values were higher for this observer, only segmentations of pituitary gland and pituitary stalk presented significant differences with respect to results provided by our approach. An example of multi-group ANOVA comparison between the four groups (three manuals and one automatic) is shown in Figure 6.21. It displays the multi-group comparison of DSC results when segmenting the chiasm. Blue indicates the group representing results from the proposed approach. Whilst in red are drawn groups with means significantly different from our method, grey indicates that results from group 1, i.e. observer 1, do not present significant differences with respect to our approach.

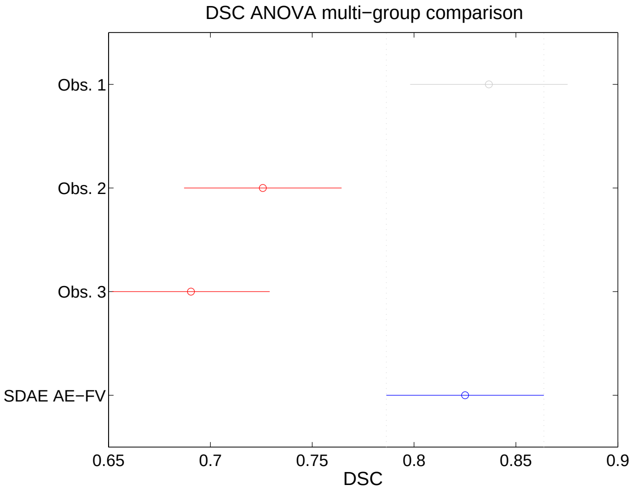


Figure 6.21: Multi-group comparison of DSC results of manual and our proposed approach for the chiasm.

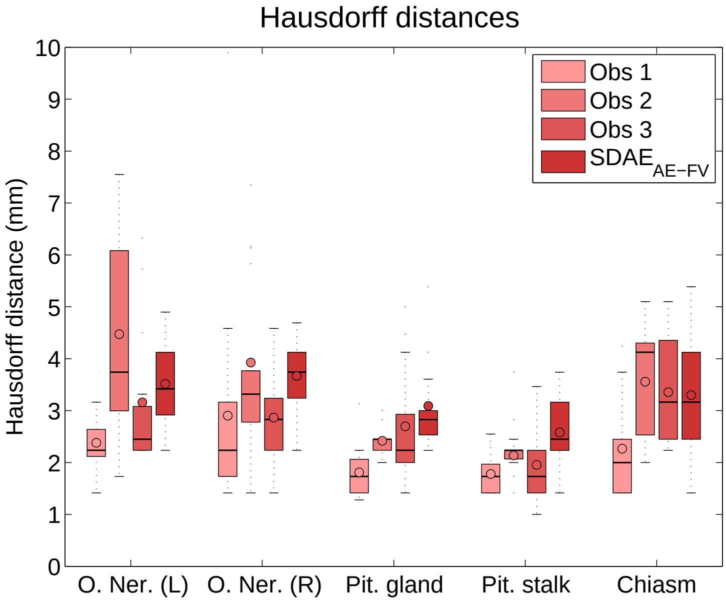


Figure 6.22: Hausdorff distance results of manual and our proposed approach for OARs of group B.

Hausdoff distances. Figure 6.22 plots Hausdorff distances distributions for the group of manual observers and the automatic proposed approach. Mean HD values for the three observers ranged from 1.78 to 4.47 mm across all the OARs. Maximum and minimum values for the automatic approach ranged from 2.58 to 3.67 mm. Mean HD values for each of the OARs for observer 1 were 2.38(± 0.47), 2.91(± 2.17), 1.81(± 0.52), 1.78(± 0.41) and 2.27(± 0.97) mm for left optic nerve, right optic nerve, pituitary gland, pituitary stalk and chiasm, respectively. Manual delineations from observer 2 provided mean HD values of 4.47(± 1.96), 3.93(± 1.89), 2.42(± 0.31), 2.14(± 0.61) and 3.56(± 1.05) mm, while mean HD values for observer 3 were 3.16(± 1.32), 2.86(± 0.85), 2.70(± 1.08), 1.96(± 0.68) and 3.35(± 0.99) mm, respectively. Finally, contours automatically generated by our approach provided the following mean HD values: 3.51(± 0.87), 3.67(± 0.67), 3.09(± 0.85), 2.78(± 0.76) and 3.29(± 1.19) mm, in the same order. Although minimum HD values were not decreased when employing the deep learning scheme, they ranged inside the variability of the experts or very close to values obtained by manual delineation. Furthermore, variability of reported HD values was decreased by the proposed system for some organs in comparison to some observers. Such is the case in both optic nerves in relation with observer 2 and 3. Variability of HD in segmenting the left optic nerve by observer 2 and 3 was of 1.96 and 1.32. Respectively, HD variability of right optic nerve was 1.89 and 0.85. By

ANOVA analysis (HD)					
	Within-subjects ANOVA	**Paired ANOVA**			
	All		Obs 1	Obs 2	Obs 3
Optic nerve L	0.0005	SDAE$_{AE-FV}$	0.0001	0.0938	0.3926
Optic nerve R	0.0014	SDAE$_{AE-FV}$	0.2015	0.6217	0.0261
Pituitary gland	0.0002	SDAE$_{AE-FV}$	0.0001	0.0077	0.2768
Pituitary stalk	0.0040	SDAE$_{AE-FV}$	0.0004	0.0659	0.0418
Chiasm	0.0064	SDAE$_{AE-FV}$	0.0142	0.5359	0.8895

Table 6.14: P-values of the ANOVA for Hausdorff distances results of the OARs of group B.

employing the proposed system this variability decreased to 0.87 and 0.66 for the left and right optic nerve.

The within-subjects ANOVA test conducted on the HD of all the groups indicated that there were significant differences among them (Table 6.14) in all the OARs. Differences on HD values were significantly important between the automatic approach and observer 1 in almost all the OARs, as reported by the paired ANOVA tests. Nevertheless, when comparing HD values from our approach with those from observer 2 and 3, differences were not significantly important in most of the cases. As example of differences statistically significant between observers 2 or 3 and our proposed approach we can find the HD results from segmentations of pituitary stalk between observer 3 and the proposed system (Figure 6.23). In addition to observer 3, differences with respect to observer 1 were also significantly important in this case. In this figure, blue represents the automatic group, while red represent the manual groups which had means significantly different.

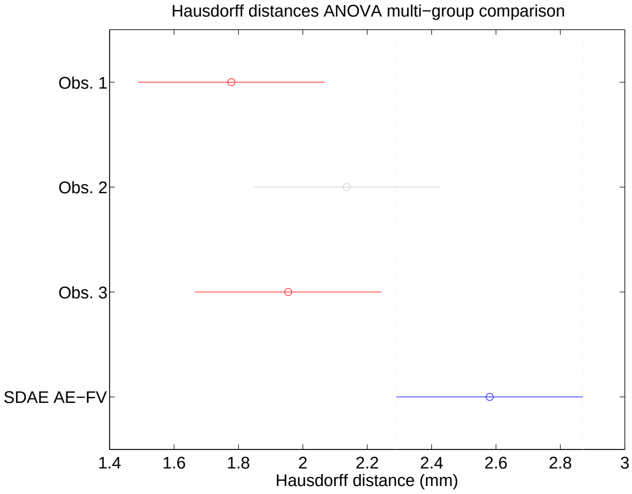


Figure 6.23: Multi-group comparison of HD results of manual and our proposed approach for the pituitary stalk.

Relative Volume differences. Distribution of relative volume differences (rVD) across all the OARs for the four groups is plotted in Figure 6.24. Segmentations from observer 1 presented the lowest rVD among the four groups, with a mean value of 11.55% (± 12.78) over all the OARs. Mean rVD over all the OARs for segmentations of observer 2 and 3 were 22.80% (± 25.24) and 18.17% (± 15.11), respectively. Last, segmentations generated by the proposed classification scheme provided a mean rVD of 17.24% (± 10.67) over all the organs. Isolating results by group and organ, segmentations from observer 1 achieved the lowest mean rVD values across all the OARs. These values were reported to be of 10.34% (± 7.01), 8.78% (± 7.94), 5.69% (± 5.28), 24.51% (± 20.42) and 8.40% (± 8.31) for left optic nerve, right optic nerve, pituitary gland, pituitary stalk and optic chiasm, respectively. For both optic nerves and pituitary stalk, contours from observer 2 obtained the highest mean rVD values, which were 26.26%, 22.78% and 26.12%, respectively. Observer 3 produced the segmentations of the chiasm with highest mean rVD values. And last, our method was ranked at last when segmenting the pituitary gland, with a mean rVD value of 18.09%.

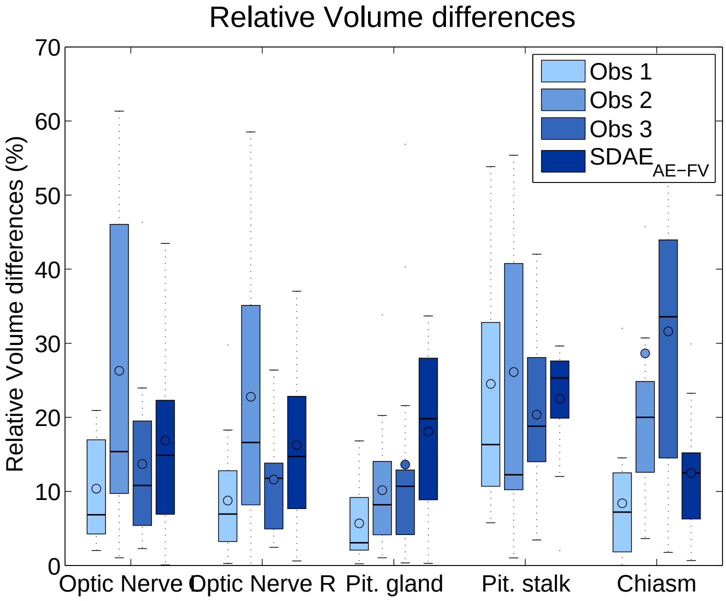


Figure 6.24: Volume differences results of manual and our proposed approach for OARs of group B.

An important point to take into consideration for the paired ANOVA analysis is that real values of relative volume differences are analyzed in these tests, instead of absolute values. Thus, results obtained by the ANOVA tests (table 6.15) may not correspond with the graphics on figure 6.24, where absolute volume differences were employed. Results extracted from volume differences

ANOVA analysis (Vol Diff)					
	Within-subjects ANOVA	**Paired ANOVA**			
	All		Obs 1	Obs 2	Obs 3
Optic nerve L	0.0004	SDAE$_{AE-FV}$	0.0001	0.0013	0.0763
Optic nerve R	0.0057	SDAE$_{AE-FV}$	0.0199	0.0069	0.0519
Pituitary gland	0.3519	SDAE$_{AE-FV}$	0.4769	0.1794	0.8961
Pituitary stalk	0.0006	SDAE$_{AE-FV}$	0.0024	0.1782	0.1295
Chiasm	0.5287	SDAE$_{AE-FV}$	0.9985	0.8655	0.1827

Table 6.15: P-values of the ANOVA for volume differences results of the OARs of group B.

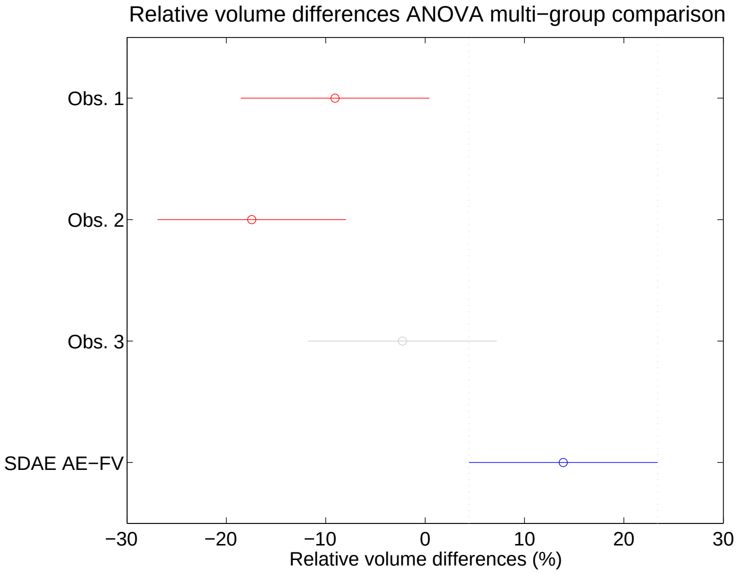


Figure 6.25: Multi-group comparison of relative volume differences results of manual and our proposed approach for the left optic nerve.

presented significant differences between groups in three out of five OARs, as indicated by the within-subjects ANOVA tests ($p < 0.05$). The paired ANOVA tests showed that rVD results were significant different between observer 1 and our approach when segmenting both optic nerves and pituitary stalk. In the same way, segmentations of both optic nerves presented significant differences, in terms of rVD, between observer 2 and our automatic approach. However, segmentations generated by our method did not show differences significantly important with respect to segmentations of observer 3. An example of ANOVA multi-group comparison is shown in Figure 6.25. The blue line represents the automatic setting. Red lines symbolize the groups comprising the manual raters which means presented statistically significant differences respect to it.

Some visual examples of manual and contours generated by our approach are shown in Figure 6.26. These images display contours of left and right optic nerves. From these images, it can be observed that automatic contours (in red) are typically between the variability of manual contours (in blue, yellow and magenta). This fact is supported by the results presented in previous section.

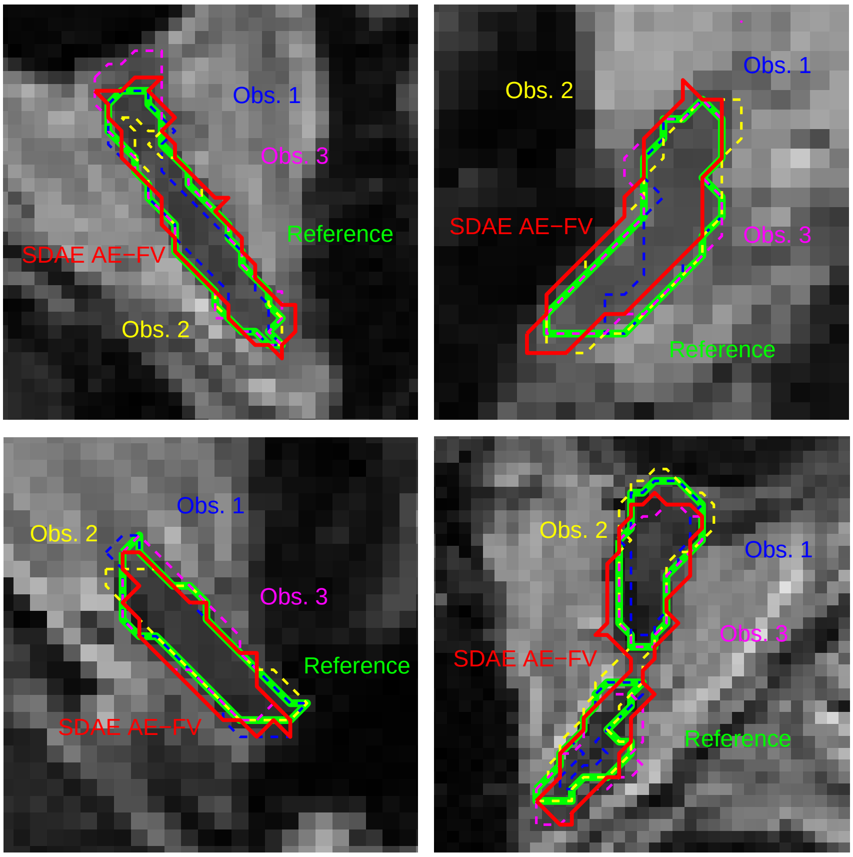


Figure 6.26: Segmentation results produced by the proposed classification system and comparison with the manual annotations.

6.2.3 Segmentation time

To compare segmentation times across all the OARs we analyze several classifier configurations. Basically, we are interested in obtaining times of SVM$_1$ and SDAE$_1$ configurations to be able to evaluate differences related to the employed classifier. In addition, time for both classifiers containing the proposed features is also evaluated, which represents the last features vector for each group. Therefore, for simplicity, we will refer to this configuration to as SVM$_{Last}$ and SDAE$_{Last}$, respectively. This is interesting to investigate whether adding more features into the classifier has repercussions on the classification time. Thus, SVM$_{Last}$ or SDAE$_{Last}$ for the OARs from group A will represent the set of enhanced features in Table 6.2. On the other hand, SVM$_{Last}$ or SDAE$_{Last}$ for OARs belonging to group B will be composed by the set of proposed features, AE-FV, which is presented in the same table.

	Segmentation time (seconds)			
	SVM$_1$	**SVM$_{Last}$**	**SDAE$_1$**	**SDAE$_{Last}$**
Brainstem	51.7023 (± 4.4485)	24.7518 (± 3.1274)	0.2460 (± 0.0145)	0.1793 (± 0.0262)
Eye (L)	12.8483 (± 0.3949)	6.2158 (± 0.2451)	0.0827 (± 0.0092)	0.0381 (± 0.0034)
Eye (R)	12.0402 (± 0.5351)	5.9896 (± 0.2185)	0.0871 (± 0.0105)	0.0374 (± 0.0031)
Lens (L)	1.1560 (± 0.1212)	1.0164 (± 0.3240)	0.02857 (± 0.0031)	0.02075 (± 0.0016)
Lens (R)	1.2104 (± 0.1564)	1.0921 (± 0.2972)	0.02913 (± 0.0033)	0.02172 (± 0.0013)
Optic nerve (L)	173.4234 (± 5.4534)	221.3296 (± 6.7034)	0.1915 (± 0.0124)	0.2628 (± 0.0172)
Optic nerve (R)	167.7524 (± 6.7484)	214.4560 (± 9.3614)	0.1726 (± 0.0091)	0.2517 (± 0.0194)
Pituitary gland	15.5368 (± 0.7802)	19.3440 (± 0.8235)	0.0536 (± 0.0066)	0.0748 (± 0.0065)
Pituitary stalk	3.0150 (± 0.1485)	4.1328 (± 0.3899)	0.0146 (± 0.0018)	0.0262 (± 0.0027)
Chiasm	5.2022 (± 0.3214)	5.8751 (± 0.5424)	0.0628 (± 0.0065)	0.1315 (± 0.0124)

Table 6.16: Segmentation times.

Table 6.16 presents mean segmentation times for first and last features sets, as explained, for both SVM and SDAE classifiers. Mean times for SVM based systems ranged from few seconds in small structures, to one or several minutes in large structures or structures presenting large shape variations. The use of proposed features into the SVM classifiers modified segmentation times. While in OARs of group A segmentation time was reduced to nearly half in most cases when employing the proposed features set, segmentation time of OARs of group B increased. This is reasonable if we take into account that sizes of proposed features sets was smaller in group A and larger in group B. On the other hand, SDAE based classification schemes achieved the segmentation in less than a second, for all the OARs. Regarding the use of proposed features, the same trend than in SVM groups is observed.

6.3 Discussion

According to some structure characteristics, results have been divided into two groups. As a reminder, group A comprises organs with homogeneous texture and small shape variation, such as the eyes or brainstem. On the other hand, organs with heterogeneous texture, and large variations in shape and locations are included in group B. For instance, optic nerves or chiasm are contained in this second group. Results from proposed system have been compared with a machine learning approach which has been widely and successfully employed for classification, i.e. support vector machines. Additionally to traditional spatial and intensity based features used in machine learning approaches, the inclusion of several features has been proposed and evaluated. These features are usually organ-dependent, and their evaluation has been performed accordingly to the division of groups A and B.

Results provided in this work demonstrate that the proposed deep learning-based classification scheme outperformed all classifier configurations taken into account in the present work. These configurations comprised, either SVM or SDAE as classifier, and one of the features sets evaluated. The basic setting in each of the classifiers was composed by classical features, i.e. spatial and intensity based features. The addition of the novel features, i.e. Geodesic transform map and LBTP-3D for OARs of group A and the AE-FV for OARs of group B, in the classifier increased volume similarity at the same time that reduced Hausdorff distances. Across all the OARs, proposed classifications schemes for groups A and B achieved the best results for similarity, surface and volume differences. Sensitivity and specificity also benefited from the use of the proposed classification scheme. First, sensitivity values were higher in SDAE based configurations than in SVM based settings. Second, the inclusion of suggested features into the classification scheme improved sensitivity values with respect to the other SDAE based settings. This trend was identified in all the OARs from both groups. Specificity values achieved by proposed systems in both groups were in around half of the cases among the top-ranked ones. Unlike in sensitivity case, specificity did not show any pattern with respect to either the classifier or the features set employed. Nevertheless, combination of higher sensitivity and specificity metrics obtained from proposed classifiers indicated that our system correctly identified more tissue voxels than the others settings did, and also was better at rejecting tissue voxels that were not related to the tissue class of interest. Statistical analysis on automatic segmentations demonstrated that results achieved by the proposed system were typically significantly different from the other groups. Particularly, significant differences often came from both SVM settings in relation with the proposed scheme. In addition, significant differences existed between SDAE settings employing classical or proposed features in some OARs.

It is important to note that similarity metrics are very sensitive in small organs. Differences in only few voxels can considerably increase or decrease comparison values. Therefore, we consider that having obtained DSC values higher than 0.7 in small OARs is very satisfactory, in addition with good values for the other metrics. Even in the worst cases, where DSC was above 0.55-0.60 for all the organs analyzed, the automatic contours can be considered as a good approximation of the reference. As example, Figure 6.27 shows the best and worst segmentation for both left and right optic nerves. While best segmentations achieved a DSC of 0.80 and 0.84 for left and right optic nerve (*top*), respectively, DSC values for worst segmentations were 0.64 and 0.60

(*bottom*).

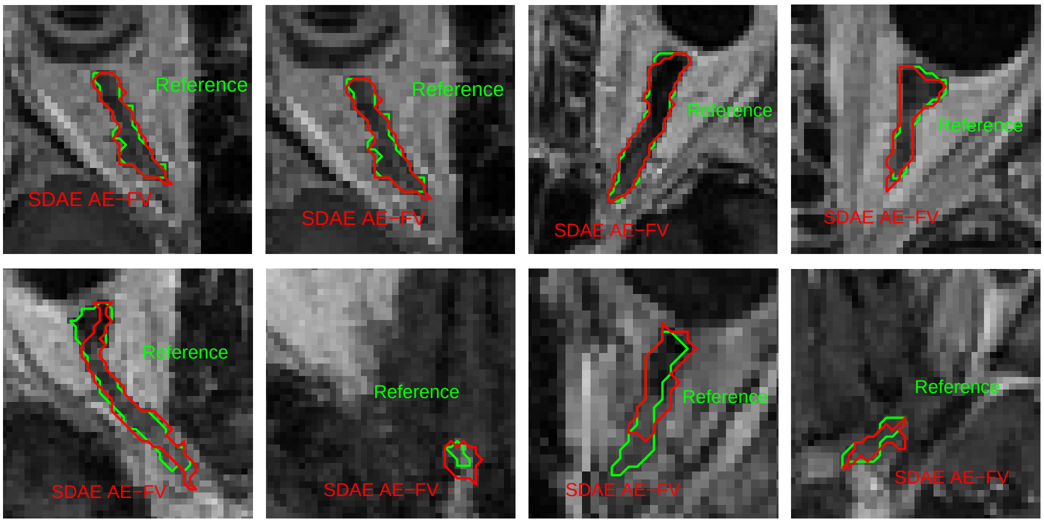


Figure 6.27: Best and worst optic nerves segmentations generated by the proposed deep learning approach. While best segmentations are shown on the top, worst segmentations cases are shown on the bottom.

In RTP context, a method that is capable of managing deformations and unexpected situations on the OARs is highly desirable. The employed dataset contained some cases where tumors inside the brainstem changed its texture properties. The proposed method correctly discarded voxels inside the brainstem that indeed belonged to tumor regions in some patients. In some others, however, tumor and brainstem were both considered as brainstem. Figure 6.28 presents a successful (*top*) and an unsuccessful (*bottom*) case. Images in the first column show segmentations generated by the four settings of group A and the reference contours. While segmentations generated by settings including proposed features, i.e SVM_2 and SDAE_2, successfully differentiate between brainstem and tumor in the top case, they included both in the segmentation of the brainstem in the bottom case. The reason of this effect mainly lies on the use of the geodesic distance transform map. This feature encourages spatial regularization and contrast-sensitivity. To generate this transformation, as it was presented, a binary mask is required. This mask is obtained in 3D from the probability map of the brainstem and it is used to seed the beginning of the geodesic map. Since this mask is eroded to ensure it will fall inside the brainstem, it will happen that the binary mask to generate the geodesic map will not appear in all the analyzed slices, particularly on both extremes. Hence, if some intensity values are not taken into account when starting the geodesic transform map, they will present differences on the

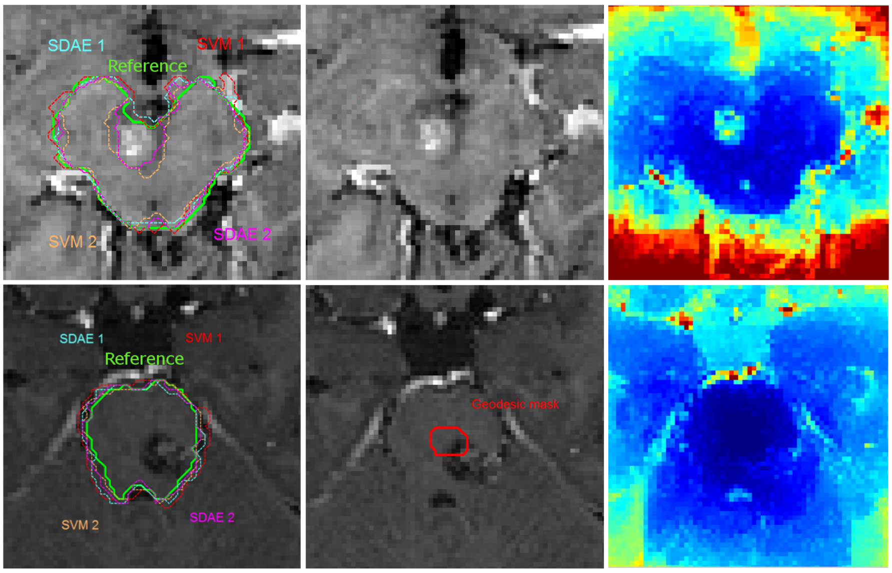


Figure 6.28: Axial slice of brainstem segmentation with tumors causing properties changes in the inner texture for two patients (*left*). Corresponding binary masks (*middle*) to generate the geodesic distance transform map (*right*) are also shown.

geodesic map. This is the case of patient shown in Fig 6.28, *top*, where tumor is located at the limit superior of the brainstem. Therefore, the geodesic map generated in this patient will make a big difference between homogeneous texture (brainstem) and heterogeneous texture (tumor), as can be seen in the *top-right* image on this figure. Contrary, patient shown in the *bottom* row presents a tumor approximately in the middle of the brainstem. In this case, binary mask employed to generate the geodesic map will contain brainstem and tumor regions. Consequently, the geodesic transform will assign similar values to these textures, since both are taken into account when creating the geodesic map (Fig 6.28), right, bottom.

In regard to comparison with manual annotations, the segmentation error we have obtained is comparable to the inter-rater difference observed when contours are delineated without time constraints. This is supported by the results obtained when comparing with manual raters. In those comparisons, we can observe that segmentation results generated by the proposed approach lie on the variability on the experts in most cases. Statistical analysis on results from manual and the proposed classification scheme point out that differences among them were not generally statistically significant. In addition, in cases where differences were significant, our automatic classifier outperformed manual rater that presented those significant differences. We can thereby say that automatic contours generated by the proposed classification system are similar to manual annotations. Therefore, its inclusion in RTP should not represent differences with respect to the use of manual contours.

Among the approaches proposed to segment brain structures included in the RTP, atlas-based techniques have attracted most attention from research. In the evaluation made by Babalola et al. [190], four approaches to segment the brainstem were compared: an atlas-based approach called Classifier Fusion and Labelling (CFL), two statistical based models - Profile Active Appearance Models (PAM) and Bayesian Appearance Models (BAM)- and an Expectation-Maximisation-based approach (EMS). CFL method provided the most accurate results, with a mean DSC of 0.94 and mean percentage rVD of 3.98 for the BS. However, segmentation time for all the OARs was reported to be 120-180 minutes. The two statistical based models - PAM and BAM - provided DSC values of 0.88 and 0.89, and percentage mean rVD of 6.8 and 7.8, respectively. Segmentation time was less than 1 min per structure for the first statistical approach and 5 min for the second one. Nevertheless, while PAM approach required a pre-registration step which took around 20 min, linear registration required by BAM took around 3 min. The last approach, EMS, underperformed the other 3 approaches, with a mean DSC of 0.83 and percentage mean rVD of 21.10, and 30 minutes to segment all the OARs involved.

Other structures, such as optic nerves and optic chiasm, have also benefit from the trend to employ atlas based approaches for segmentation. An atlas-navigated optimal medial axis and deformable model algorithm (NOMAD)to segment these two structures in MRI and CT images was presented in the work of Noble and Dawant [79]. Ten CT/MRI pairs were used for evaluation purposes. Mean DSC values achieved for the testing set were just below 0.8 for both the optic nerves and at 0.8 for the chiasm. In their work, they also reported that segmentation error obtained was comparable to the inter-rater difference observed when contours were delineated without time constraint in a laboratory setting. Segmentation of the optic nerves in a test volume required approximately 20 minutes. As alternative to atlas-based methods, Bekes et al. [124] proposed a geometrical model-based segmentation technique. In addition to optic chiasm and nerves, the eyes and lenses were also included in the evaluation, where sensitivity and specificity are used instead of DSC. Mean sensitivity values of 97.71, 97.81, 65.20 and 76.79 were achieved by their method when segmenting the eyes, lenses, optic chiasm and optic nerves, respectively. Analogously, reported mean specificity values were 98.16, 98.27, 93.50 and 99.06. The running time for all the structures was around 6-7 seconds for a whole CT volume. Whilst segmentation of eyes and lenses were satisfactory, segmentation of optic nerves and chiasm was below their expectations. Repeatability and reproducibility of the automatic results made the method not being usable for RTP for these two challenging structures.

Additionally to the presented works, where segmentation is done in healthy patients, other works have focused on the evaluation of segmentation performance of one or a set of OARs in radiotherapy context. Bondiau et al. [5] presented a study aiming to evaluate an atlas-based method to segment the brainstem in a clinical context. To carry out such evaluation, a total of 7 experts and 6 patients were employed. The automatic method achieved a mean sensitivity value of 0.76, which was below the mean sensitivity of any of the experts. However, only in 2 out of the 6 cases the automatic approach presented the lowest sensitivity value. In the other four cases, sensitivity was between the expert variation. With regards to the specificity, means of the experts ranged from 0.86 to 0.99, whilst it was 0.97 for the automatic approach. Volume measurements revealed that, although the automatic results mostly lie between the variability of the experts, it tend to underestimate the segmented volume with respect to the mean of the manual delineations. With these results, authors suggested that this method provided a good trade-off between accuracy and robustness. Additionally, reported results could be comparable to those from the experts. Results reported that the total duration of the automatic segmentation process to obtain a fully labeled MRI was of 20 min.

In the work of Isambert et al. [78] another atlas-based segmentation (ABAS) software, which is included in a commercial solution for RTP, was also evaluated in therapy clinical context. Automatic segmentations of the brainstem, cerebellum, eyes, optic nerves, optic chiasm and pituitary gland of 11 patients on MRI T1-weighted images were evaluated. It was found that for large organs, DSC reported values were higher than 0.8; whereas for smaller structures, DSC was lower than 0.4. More specifically, mean DSC distribution across all the OARs was: 0.85, 0.84, 0.81, 0.38, 0.41 and 0.30, for the brainstem, cerebellum, eyes, optic nerves, optic chiasm and pituitary gland, respectively. With exception of the optic nerves, the atlas-based approach underestimated all the volumes from 15 % in the case of the brainstem to 50 % when segmenting the optic chiasm. The mean time required to automatically delineate the set of 6 structures was 7-8 min. Following the ROC analysis that we also employed in the present work, segmentations generated by the automatic approach were clinically acceptable for the brainstem, eyes and cerebellum. On the other hand, all the segmentations for the optic chiasm, and most of the segmentations for optic nerves and pituitary gland were considered as poor.

In a more recent study on RTP context, Deeley et al. [6] compared manual and automated approaches to segment brain structures in the presence of space-occupying lesions. The objective of this work was to characterize expert variation when segmenting OARs in brain cancer, and to assess an au-

Structure	Ref	Image Modality	Method	DSC	HD(mm)	Vol Dif (%)	Sens	Spec	Time
Brainstem	Babalola [190]	MR	Atlas	0.94	4.83	3.98	-	-	120-180 min.[1]
			Statistical (PAM)	0.88	6.02	6.80	-	-	1 min[2]
			Statistical (BAM)	0.89	6.37	7.80	-	-	5 min.[3]
			Expectation-minimization	0.83	7.74	21.10	-	-	30 min.[1]
	Bondiau [5]	MR	Atlas	-	-	-	76.17	96.67	20 min.
	Deeley [6]	MR	Atlas	0.86	-	-	-	-	-
	Fritscher [191]	CT	Atlas	0.86	8.00	-	-	-	-
	Hoang [192]	CT	Atlas	0.84	3.4	-	-	-	21 sec.[4]
	Isambert [78]	MR	Atlas	0.85	-	-14.8	79.40	96.89	7-8 min[5]
	Our method	MR	Deep Learning	0.92	5.87	3.10	90.56	91.43	0.1793[6]
Eyes	Bekes [124]	CT	Geometrical	-	-	-	97.71	98.16	1 sec.
	Deeley [6]	MR	Atlas	0.84	-	-	-	-	-
	Hoang [192]	CT	Atlas	0.59	5.9	-	-	-	12 sec.[4]
	Isambert [78]	MR	Atlas	0.82	-	26.26	70.33	98.36	7-8 min[5]
	Our method	MR	Deep Learning	0.89	5.15	10.08	90.92	98.94	0.0378 sec.[6]
Lenses	Bekes [124]	CT	Geometrical	-	-	-	97.81	98.27	1 sec.
	Our method	MR	Deep Learning	0.76	2.17	22.81	87.23	75.22	0.0212[6]
Pituitary gland	Isambert [78]	MR	Atlas	0.3	-	-36.8	23.68	82.91	7-8 min[5]
	Our method	MR	Deep Learning	0.76	3.33	18.10	84.22	82.69	0.0748 sec.[6]
Optic chiasm	Bekes [124]	CT	Geometrical	-	-	-	65.20	93.50	1 sec.
	Deeley [6]	MR	Atlas	0.37	-	-	-	-	-
	Hoang [192]	CT	Atlas	0.57	5.80	-	-	-	10 sec.[4]
	Isambert [78]	MR	Atlas	0.42	3.10	-50.5	30.36	93.27	7-8 min[5]
	Noble [79]	CT	Atlas	0.8	-	-	-	-	-
	Our method	MR	Deep Learning	0.83	3.30	12.48	83.94	86.11	0.1315 sec.[6]
		MR							
Optic nerves	Bekes [124]	CT	Geometrical	-	-	-	76.79	99.06	1 sec.
	Deeley [6]	MR	Atlas	0.52	-	-	-	-	-
	Harrigan [193]	CT	Atlas	0.77	3.75	-	-	-	-
	Isambert [78]	MR	Atlas	0.38	-	31.3	37.91	76.86	7-8 min[5]
	Noble [79]	CT & MR	Atlas	0.79	4.20	-	-	-	20 min.
	Panda [81]	CT	Atlas	0.77	3.33	-	-	-	-
	Our method	MR	Deep Learning	0.79	3.59	16.56	81.89	89.56	0.2572[6]

[1] Set of brain structures
[2] It required a pre-registration step which took 20 min.
[3] It required a pre-registration step which took 3 min.
[4] It required a pre-registration step which took 50 min.
[5] Set of 6 OARs
[6] It requires between 1-6 seconds to extract features, depending on the structure.

Table 6.17: Summary of segmentation results of related works.

Reference	Number of patients	Reference contours	Notes
Babalola [190]	270 Images	N.A.	
Bekes [124]	41 Images	N.A.	Not gold standard. STAPLE
Bondiau [5]	20 Images (Training) 6 Images (Testing)	7 physicians	No lesions inside the brainstem
Deeley [6]	20 Images	8 experts	Large space-occupying lesions. Often close to the OARs
Fritscher [191]	18 Images	1 expert	
Harrigan [193]	501 images from 183 patients	N.A.	
Hoang [192]	100 patients	2 experts	Expert 1 contoured 43 images. Expert 2 contoured 57 images.
Isambert [78]	11 patients	2 experts	The two experts made the contours together.
Noble [79]	4 Images (Model training) 10 Images (Parameter training) 10 Images (Testing)	1 observer	A student made the contours. Then, corrected by 2 experts.
Panda [81]	30 patients	1 expert	

Table 6.18: Experimental setting-up of related works.

tomatic segmentation method in such context. To achieve the automation of the segmentation process, a registration-driven atlas-based algorithm was employed. A set comprising the brainstem, optic chiasm, eyes and optic nerves was evaluated. Main results disclosed in their evaluation showed that the analyzed automatic approach exhibited mean DSC values between 0.8-0.85 for larger structures, i.e. brainstem and eyes. Contrary, DSC reported for smaller structures, i.e. optic chiasm and optic nerves, were of 0.4 and 0.5, respectively. Results demonstrated that although both manual and automatic methods generated contours of similar volumes, experts exhibited higher variation with respect to tubular structures. Coefficients of variation across all the patients ranged from 21-93 % of mean structure volume.

Although presented works have demonstrated to perform well when segmenting some structures, most of them have resulted ineffective when applied to a multi-structure environment. In this context, this situation is aggravated if small structures, such as the chiasm, are included in the segmentation. On the other hand, works presenting the highest results represented the longest ones in terms of processing times. These times ranged from 1 or 2 minutes to several minutes, per structure. A summary of the performance from these previous works, as well as from our proposed method, are presented in Table 6.17. In this table we observe that, in terms of similarity metrics (volume and surface), our method beat all other works in most situations. Additionally, a noteworthy aspect to highlight from our approach is its significantly low segmentation time, which is several orders of magnitude in comparison with the others. In order to have a more relevant comparison between methods, table 6.18 is added. In this table, experimental settings up for each work shown in table 6.17 are presented. With all this we may thereby say that the presented approach outperforms, up to date, to all the other segmentation methods to segment OARs in brain cancer.

Results also demonstrate that the proposed deep learning-based classification scheme outperformed all previous works when segmenting the set of OARs analyzed. Nevertheless, it is important to note that differences in data acquisition, as well as metrics used to evaluate the segmentation, often compromise comparison to other works. Although it was not possible in this work to use the same datasets as those used in previous studies, the consistently higher performance our approach achieved, as indicated by the results, suggests that the method presented in this thesis outperforms previously presented approaches to segment these structures. Results show that by employing SDAE as classifier, segmentation time was significantly reduced in comparison to other classical machine learning approaches, such as SVM. This is particularly noteworthy if we take into consideration that most works referenced in this thesis to segment involved structures are atlas-based, and therefore registration dependent. This makes their segmentation times very expensive in comparison with the proposed approach, which is between two and three orders of magnitude faster. Current implementation of the proposed system is not computationally optimized and the bottle neck of the process is the features extraction step, which processing time ranges between 1-6 seconds for each of the OARs. Although it is not an expensive stage, it represents more than 95% of the total process. Since the extraction of the features does not require difficult programming operations, its parallelization is easily affordable. This may substantially decrease the whole segmentation process up to segmentation times below one second for an entire organ.

One of the strengths of deep learning methods relies on their ability to transfer knowledge from human to machine. They 'learn' from a given training data set. Hence, for example, when no visible boundaries are present, the classifier uses its transferred intelligence from doctors to perform the segmentation as they would do. As a prove of this learning, results presented in this thesis have shown how well the proposed system learned from the available dataset.

Nevertheless, one of the main concerns of this thesis was the generation of a simulated ground truth. It was obtained in this work by using the computationally simple concept of probability maps. In this method, which is analogous to the voting rule approach, probability maps were thresholded at a variable level in order to create the mask. Although thresholds values were fixed according to the number of available contours, which also corresponded with the suggestion of Biancardi [185], thresholding probability maps at a static predetermined level may be problematic. Determination of the most suitable threshold for each organ presents a challenge. A reasonable first choice is to fix its value to 50% as it represents the threshold for majority voting rule. Nevertheless, as pointed out in the work of Deeley [6], 50% might not be reliable with such statistically small number of raters with unknown individual variance. Thus, an appropriate threshold value for one cohort of experts may not suit for another different cohort. The same reasoning can be extended for different organs, where consensus among raters is dependent on organ. Therefore, to be able of simulating more consistent reference standard we encourage further studies to involve more experts in the manual delineation step.

Chapter 7
Conclusions and Future Work

"The two most important days in your life are the day you are born and the day you find out why."
Mark Twain

In this chapter we review the motivations for this work, as well as the important contributions of this thesis. Following this, possible future directions are also discussed.

7.1 Discussion

This dissertation addresses the problem of organs at risk segmentation in brain cancer towards enabling its adoption in clinical routine. To achieve this, the work in this thesis puts forth a practical application in the field of medical image segmentation of one of the hottest research topics nowadays, i.e. deep learning.

Segmentation of medical images is a field that have spurred an overwhelming amount of research. However, open issues abound with regard to approaches to segment organs at risk in brain cancer and its usability in clinical practice. Nowadays, and up until a few years ago, atlas and statistical based models have represented the most employed techniques to perform a sort of automatic delineation in medical images, particularly for brain structures. However, they present some disadvantages and therefore suffer from slow adoption in clinical routine.

Atlas-based segmentation approaches rely on registration techniques. In these methods, anatomical information is exploited by means of images already annotated by experts, referred to as atlases, to be matched on the patient under examination. To compute such transformation, deformable registration is often used. After registration, deformed contours are transferred to the target image. The quality of the deformed contours directly depends on the quality of the deformation. Nevertheless, this is difficult to evaluate. Furthermore, registration techniques encompasses regularization models of the deformation field, whose parameters are complex to adjust, particularly in inter-patient cases. Another limitation of atlas-based methods is that contours included in the atlases contain prior knowledge about organs pictured in the image which is not commonly exploited. To perform the segmentation, contours are merely deformed. As a consequence, most of the information conveyed by the contours, such as shape or appearance, remains implicit and likely underexploited. Statistical models present an alternative to address this issue by making a more explicit use of such prior information to assist the image segmentation. Unlike atlases, images are not registered but shapes and, sometimes, the appearance of the organ, are learned in order to be found in a target image. Because of target points are searched in a local constrained vicinity of the current estimation for each location, a sufficiently accurate initialization needs to be provided to make the model converge to the proper shape. Therefore, search of shape and/or appearance requires an initialization. If the initial position is too distant from the searched object, in terms of translation, rotation or scale, this can lead to poor object identification. Details of these, and other published works to segment brain structures were disclosed in Chapter 3.

The objective of this thesis was therefore to propose an approach as alternative to these existing methods that also addresses their limitations. Particularly, an organ segmentation scheme based on a deep learning technique was suggested. This approach, as most of machine learning based methods, attempts to reproduce the way radiation oncologists manually delineate the organs. First, all information required to learn how to segment each of the organs at risk is extracted from images where organs were delineated. This information is transformed into a features array that serves as input of the network. Then, the network learns a hierarchical representation of the input, which is later employed for classification. The strength of deep architectures is to stack multiple layers of nonlinear processing, a process which is well suited to capture highly varying functions with a compact set of parameters. The deep learning scheme, based on a greedy layer-wise unsupervised pre-training, allows to position deep networks in a parameter space region where the supervised fine-tuning avoids local minima. Deep learning methods achieve very good accuracy, often the best one, for tasks where a large set of data is available, even if only a small number of instances are labeled. In addition, deformable registration techniques are no longer required in our approach, but a simple manual rigid alignment. Even though we have not investigated a solution to automatically perform the required alignment, it should be easily automatized.

Details of the technique employed to create the deep network, as well as the features propose to improve the segmentation performance were introduced in Chapter 4, where main contributions of this work were presented. The learning network is generated by stacking denoising auto-encoders in a deep architecture. To train a learning system, a set of features is commonly fed into the network. Particularly, in deep learning architectures, such as convolutional neural networks, restricted Boltzman machines, or even auto-encoders, two or three-dimensional patches from one or multiple images are typically employed as features vector. From these patches, the network unsupervisedly learns the most discriminative representation of the input. Although they have demonstrated to break records in several domains, such as speech recognition or image classification, we consider that by using patches they do not fully exploit relevant information coming from traditional machine learning approaches to analyze medical images in general. This unexploited knowledge may come in the form of likelihood to belong to some class, voxel location or textual properties, for example. Thus, the feature set proposed in this work is very different from features sets employed on most of the deep learning settings applied to medical imaging.

Typical machine learning schemes to segment medical images employ pixel or voxel intensity values, intensity of a neighboring area, likelihood of belonging to a given class and location. Although these hand-crafted features may be sufficient for some well-defined structures, they do not provide the best solution when attempting to segment challenging structures. Thus, additional features have been proposed in this thesis to enhance the discriminative power of the features vector. Since properties are different from one organ to another, some of the proposed features are organ-dependent. Hence, for example, for organs with homogeneous texture and small shape variations we have proposed features that encourage spatial regularization, such as the geodesic distance transform map. On the other hand, for organs with strong variations on shape and intensity we have suggested the combined use of contextual and textural properties. As a consequence, features set varies from one group of organs to another.

We designed an evaluation study to evaluate the performance of the proposed approach, quantify variation among experts in segmenting organs at risk in brain cancer, and assess the proposed automatic classification scheme in this context. First, a reference standard was created from the manual contours, which served as ground truth to compare with. To evaluate the performance of our approach, results were compared to those provided by a state-of-the-art machine learning classifier, i.e. support vector machine (SVM). In the second part of the evaluation, automatic contours generated by the proposed approach were also compared to manual annotated contours by experts.

Results demonstrated that by only employing a network composed by a stacked of denoised auto-encoders, segmentation performance increased with respect to SVM. Additionally, when proposed features were included in the features set, reported results showed that improvement on segmentation performance was noticeable. Across the experiments we noticed that segmentations of OARs of group B, in some patients, were highly different when using either augmented or textural features. While in some patients the features set composed by augmented features achieved the best results, in some other patients the best result was obtained by the textural features set. Nevertheless, when combining both of them, results were more homogeneous, which can be observed in the standard deviation on the results section. For the other groups, the classification performed with the proposed features set outperformed the classical set in most cases.

Even though the presented work is not pioneering on the evaluation of automatic segmentation of OARs in the context of brain radiation therapy, it presents important improvements respect to the others (See Table 6.17). Large structures, such as eyes or brainstem, have been successfully segmented in all previous works evaluating segmentation performance in brain cancer context. Contrary, segmentation of small structures was not always satisfactory. By employing the proposed classification system we: i) improved segmentation performance of structures already successfully segmented and ii) provided a satisfactory segmentation for those structures which segmentation could not be always achieved. Furthermore, all presented works to analyze OARs segmentation in radiotherapy context are based on atlas and thus registration dependent. This makes segmentation times to be over several minutes, which might be clinically impractical in some situations. In addition to the segmentation times, other disadvantages of atlas-based methods have already been discussed. Our method, however, performs the segmentation in few seconds for each single OAR. A noteworthy point is that features extraction represented nearly 97.5% of the whole segmentation process. Since this stage is composed by simple and independent image processing steps, this can be easily parallelized. By doing this, the total segmentation time may be drastically reduced to less than a second per structure. Another remarkable difference with respect to some other approaches is that the proposed system does not require combination of more than one image modalities.

When comparing the results with the manual contours, it can be observed that they lie inside the variability of the observes. Statistical tests demonstrated that there were not significant differences between automatic and manual delineations for many of the cases. All this, together with the remarkably low segmentation time reported in the experiments, makes this technique suitable for being used in clinical routine. Therefore, the introduction of such technique may help radiation oncologists to save time during the RTP, as well as reducing variability in OAR delineation.

This thesis has represented therefore a first step in developing and exploring deep denoised autoencoders for being applied to the segmentation of organs at risk on brain cancer. Its evaluation has been assessed in a multi-rater context. In addition, it does so without being subject to fatigue or inattentiveness, which can affect human measurements and diminish reliability in studies of large samples over a long time.

7.2 Future work

In this thesis, we have proposed an approach that solely employs information extracted from magnetic resonance images (MRI). More specifically, only the sequence T1 from the MRI set is used. Nevertheless, having employed exclusively T1 sequences might have underestimated the power of our approach. The reason is that contouring of some OARs on FLAIR or T2 sequences would probably have improved the inter-observer reproducibility without degrading learning and automatic segmentation. However, all the sequences were not available in all the patients contained on the employed dataset. We have also shown that for training and classifying we utilize the features vector. Including additional information on this vector is straightforward. Since more MRI sequences other than T1 are typically acquired to plan the treatment and diagnosis, we suggest to combine MRI-T1 with other modalities, such as T2 for example, when available. The combination of different MR sequences can enhance the segmentation, particularly on those regions where these image sequences are complimentary. Independently on the sequence added, any relevant information included into the classifier may help to improve the segmentation performance. We therefore encourage future research on this topic to include other image sequences in both contouring and learning/classification steps. Another main direction for future research is to examine the contribution of other image properties as features during the training and classification.

In this work, good segmentation performance has been reported by the proposed classification scheme by training huge networks with a relatively small amount of data. Indeed, these networks were sometimes composed by several millions of parameters, while number of training samples were of several thousands. Even though trained deep networks overfit training dataset, they still generalized pretty well to unseen samples. This may be explained by the fact that brain MRI images are highly structured, often presenting small variability between regions from one brain to another. Using more patient cases in the training set aiming to capture more variability would ideally be the best solution to prevent from overfiting. Additionally, this increase of the training set might also positively impact on classification performance. Unfortunately, labeled datasets are rare and difficult to obtain. Consequently, generation of artificial MRI cases from existing ones should be considered in further works, i.e. data augmentation. This could include small transformations such as rotations, scaling, noise or some other small distortions.

In addition, in our experiments, and in most of proposed works employing deep architectures, the number of hidden units in each layer, as well as the number of layers is manually determined. Therefore, the network architectures employed might not be necessarily optimal. By employing deep learning a better representation of an input is automatically extracted. However, network architecture is still manually tuned. I believe that performing more intensive studies such as learning optimal network structure from input data for its practical use in clinical setting would bring more power to neural networks.

The work presented in this thesis has been mostly developed within an enterprise, in an industrial environment. As such, one of the main goals of the company is to integrate this work into its products. The code developed represents a first functional prototype that can be employed to obtain results such as the ones presented in this dissertation. Nevertheless, its use in clinical routine in its current state still requires some efforts from the development side. Development of an optimized prototype would represent one of the first tasks to carry out. Although its current performance allows to segment a structure in relatively small amount of time, this process can still be optimized by programming the features extraction step on GPU. In addition to processing time, user experience is of high importance when trying to develop software that will be employed by non-experts users through an interface. Before deploying a clinical usable version of the final product, a clinical validation with a larger dataset should be also envisaged.

Chapter 8
Own Publications

[1] J. Dolz, N. Betrouni, D. Kharroubi, L. Massoptier, M. Vermandel. "A deep learning classification scheme based on augmented-enhanced features to segment organs at risk in the optic region in brain cancer". *Submitted to Journal of Medical Image Analysis by December, 14th, 2015. Major revision at 6th March, 2016.*

[2] J. Dolz, N. Betrouni, M. Quidet, D. Kharroubi, H.A. Leroy, N. Reyns, L. Massoptier, M. Vermandel. "Stacking denoising auto-encoders in a deep network to segment the brainstem on MRI in brain cancer patients: a clinical study". *International Journal of Computerized Medical Imaging and Graphics. 52 (2016) 8-18.* DOI: 10.1016/j.compmedimag.2016.03.003

[3] A. Laruelo*, J. Dolz*, S. Ken, L. Chaari, M. Vermandel, L. Massoptier, A. Laprie. "Probability map prediction of relapse areas in glioblastoma patients using multi-parametric MR", *ESTRO 35th Meeting, April-May 2016, Turin.* ***Nominated to the Best Poster ESTRO award in the category of Physics.***

[4] J. Dolz, A. Laprie, S. Ken, H.A. Leroy, N. Reyns, L. Massoptier, M. Vermandel. Supervised machine learning-based classification scheme to segment the brainstem on MRI in multicenter brain tumor treatment context. *International Journal of Computer Assisted Radiology and Surgery (IJCARS), 2015, 1-9.*

[5] J. Dolz, S. Ken, H.A. Leroy, N. Reyns, A. Laprie, L. Massoptier, M. Vermandel. Supervised machine learning method to segment the brainstem on MRI in multicenter brain tumor treatment context. *Computed Assisted Radiology and Surgery (CARS),* Barcelone, June, 2015.

[6] J. Dolz, L. Massoptier, M. Vermandel. Segmentation algorithms of subcortical brain structures on MRI for radiotherapy and radiosurgery: a survey. *International Journal of Innovation and Research in Biomedical Engineering (IRBM). 36,200-212. (2015).*

[7] J. Dolz, H.A. Leroy, N. Reyns, L. Massoptier, M. Vermandel. A fast and fully automated approach to segment optic nerves on MRI and its application to radiosurgery. *IEEE International Symposium on Biomedical Imaging (ISBI),* New York, April, 2015. (pp. 1102-1105)

[8] J. Dolz, H.A. Kirisli, M. Vermandel, L. Massoptier. Subcortical structures segmentation on MRI using suuport vector machines. In Multimodal imaging towards individualized radiotherapy treatments, pages 24-31. ISBN 978-94-6186-309-6, 2014.

Chapter 9
French Summary

De nos jours, les techniques de segmentation automatique sont rarement utilisées en routine clinique. Le cas échéant, elles s'appuient sur des étapes préalables de recalages d'images. Ces techniques sont basées sur l'exploitation d'informations anatomiques annotées en amont par des experts sur un "patient type". Ces données annotées sont communément appelées "Atlas" et sont déformées afin de se conformer à la morphologie du patient en vue de l'extraction des contours par appariement des zones d'intérêt. La qualité des contours obtenus dépend directement de la qualité de l'algorithme de recalage. Néanmoins, ces techniques de recalage intègrent des modèles de régularisation du champ de déformations dont les paramètres restent complexes à régler et la qualité difficile à évaluer. L'intégration d'outils d'assistance à la délinéation reste donc aujourd'hui un enjeu important pour l'amélioration de la pratique clinique.

L'objectif principal de cette thèse est de fournir aux spécialistes médicaux (radiothérapeute, neurochirurgien, radiologue) des outils automatiques pour segmenter les organes à risque des patients bénéficiant d'une prise en charge de tumeurs cérébrales par radiochirurgie ou radiothérapie. Pour réaliser cet objectif, les principales contributions de cette thèse sont présentées sur deux axes principaux. Tout d'abord, nous considérons l'un des derniers sujets d'actualité dans l'intelligence artificielle pour résoudre le problème de la segmentation, à savoir le "deep learning". Cet ensemble de techniques présente des avantages par rapport aux méthodes d'apprentissage statistiques classiques (Machine Learning en anglais). Le deuxième axe est dédié à l'étude des caractéristiques d'images utilisées pour la segmentation (principalement les textures et informations contextuelles des images IRM). Ces caractéristiques, absentes des méthodes classiques d'apprentissage statistique pour la segmentation des organes à risque, conduisent à des améliorations significatives des performances de segmentation. Nous proposons donc l'inclusion de ces fonctionnalités dans un algorithme de réseau de neurone profond (deep learning en anglais) pour segmenter les organes à risque du cerveau.

Nous démontrons dans ce travail la possibilité d'utiliser un tel système de classification basée sur techniques de "deep learning" pour ce problème particulier. Finalement, la méthodologie développée conduit à des performances accrues tant sur le plan de la précision que de l'efficacité.

9.1 Introduction

Le cancer représente un groupe de maladies communes, non transmissibles, chroniques et potentiellement mortelles affectant la plupart des familles dans les pays développés, et un contributeur de plus en plus important à une mort prématurée au sein de la population de ces pays [2]. En particulier, les tumeurs cérébrales sont la deuxième cause la plus fréquente de décès par cancer chez les hommes âgés de 20 à 39 ans et la cinquième cause la plus courante de cancer chez les femmes âgées de 20 à 39 ans [4].

Une tumeur cérébrale est toute masse provoquée par une croissance anormale ou incontrôlée de cellules qui surviennent à l'intérieur ou à proximité du cerveau. En général, ces tumeurs sont classées en fonction de plusieurs facteurs, y compris l'emplacement, le type de cellules impliquées, et le taux de croissance. Les tumeurs cérébrales dites "primaires" sont des tumeurs à croissance plus ou mois rapide, localisées dans le parenchyme cérébrale et sans capacité de ce propager sur des sites distants. Leur degré de croissance constitue notamment un facteur de malignité et elles sont ainsi classées "benignes" (ex. neurinomes, méngiomes) ou "malignes" (ex. gliome de bas grade, glioblastome). Les tumeurs issues d'une localisation distante (ex. poumon, foie, sein) ont une croissance généralement plus rapide et sont dites "secondaires". Il s'agit de métastases cérébrales consécutives à un cancer d'une localisation extracérébrale. Ces dernières sont toujours des tumeurs malignes. Cependant, primaire ou secondaires, bénigne ou maligne, les tumeurs cérébrales restent toujours potentiellement invalidantes et critiques pour la survie du patient.

La radiothérapie (RT) et la radiochirurgie (SRS) sont parmi l'arsenal de techniques disponibles pour traiter les tumeurs cérébrales. Le terme radiothérapie décrit les applications médicales des rayonnements ionisants pour détruire les cellules malignes en endommageant leur ADN [9]. La RT est souvent organisé en deux phases: la planification et la délivrance. Les images sont acquises, les régions d'intérêt sont identifiées et la balistique est planifiée à partir de ces données d'imagerie. Le traitement planifié est ensuite délivré au patient. Afin de calculer la dose a délivrer, la position des volumes cibles doit être précisément déterminée.

Un objectif majeur de RT est de priver les cellules cancéreuses de leur potentiel de multiplication et éventuellement tuer les cellules cancéreuses. Cependant, le rayonnement créée également des lésions aux tissus sains. Par conséquent, l'objectif principal de la radiothérapie est délivrer une dose importante à la tumeur, tout en veillant à ce que les tissus sains avoisinant soient épargnés autant que possible. En particulier pour les traitements radiochirurgicaux, où la dose de rayonnement est considérablement plus élevée et délivrée en séance unique, des erreurs de configuration ou de localisation peuvent entraîner une surdose sévère du tissu sain adjacent. Cette surexposition aux rayonnements peut conduire à des complications sévères, progressives et irréversibles, qui se produisent souvent des mois ou des années après le traitement. Ces structures critiques à conserver sont désignées comme organes à risque (OAR). Dans la RT cérébrale, les nerfs optiques, le chiasma, le tronc cérébral, les yeux, le cristallin, l'hippocampe et l'hypophyse sont généralement considérés comme OARs.

Au cours des dernières décennies, l'imagerie médicale, initialement utilisée pour la visualisation des structures anatomiques, a évolué pour devenir un outil essentiel au diagnostic, au traitement et au suivi de l'évolution des pathologies. En particulier, dans l'oncologie, l'évolution des techniques d'imagerie a permis d'améliorer la compréhension du cancer, de son diagnostic à sa prise en charge thérapeutique et du suivi évolutif. Les techniques d'imagerie médicale avancées sont donc utilisées pour la chirurgie et pour la radiothérapie. Il existe un large éventail de modalités d'imagerie médicale. Les premières méthodes d'imagerie, invasives et parfois risquées, ont depuis été abandonnées en faveur de modalités non-invasives, de haute résolution, telles que le scanner (CT) ou, en particulier, l'imagerie par résonance magnétique (IRM). L'IRM possède une sensibilité plus élevée pour détecter une tumeur, ou des changements au son sein et un meilleure contraste pour délimiter les structures cérébrales saines. Pour ces raisons, et parce que l'IRM ne repose pas sur des rayonnements ionisants, l'IRM a progressivement supplanté le CT comme pilier de l'imagerie en neuro-oncologie clinique, devenant la modalité de référence pour le diagnostic, le suivi et la planification des traitements de lésions cérébrales [26].

Parce que RT et SRS s'appuient sur une irradiation importante, la tumeur et les tissus sains environnants doivent être précisément définies. En particulier pour les OARs pour lesquels la connaissance de leur localisation et de leur forme est nécessaires pour évaluer et limiter le risque de toxicité sévère. Parmi les modalités d'image disponibles, les images IRM sont largement utilisées pour segmenter la plupart des OARs. Dans la pratique, cette délinéation est principalement réalisée manuellement par des experts avec éventuellement un faible support informatique d'aide à la segmentation [28]. Il en découle que le processus est fastidieux et particulièrement chronophage avec une variabilité inter ou intra observateur significative. Une part importante du temps médical s'avère donc nécessaire à la segmentation de ces images médicales. Si en automatisant le processus, il devient possible d'obtenir un ensemble plus reproductible des contours acceptés par la majorité des oncologues, cela permet d'améliorer la planification et donc la qualité du traitement. En outre, toute méthode de réduction du temps nécessaire à cette étape contribue à une une utilisation plus efficace des compétences de l'oncologue.

Pour remédier à ces problématiques, divers systèmes assistés par ordinateur pour (semi-) automatiquement segmenter les OARs ont été proposés et publiés au cours des dernières années. Néanmoins, la segmentation (semi-)automatique des structures cérébrales reste encore difficile, en l'absence de solution générale et unique. De plus, en raison de l'augmentation du nombre de patients à traiter, les OARs ne peuvent pas toujours être segmentés avec précision, ce qui peut conduire à des plans sous-optimaux [32]. Cela rend l'implémentation en routine clinique d'un outil de segmentation des OARs assistée par ordinateur hautement souhaitable.

9.2 Etat de l'art

La segmentation d'image est un problème de partitionnement d'une image d'une manière sémantiquement résolue. La subdivision de l'image en régions significatives permet une représentation compacte et plus facile de l'image. L'agrégation des pixels d'une forme donnée se fait selon un critère prédéfini. Ce critère peut être basé sur de nombreux facteurs, tels que l'intensité, la couleur ou la texture, la continuité des pixels, et certaines autres connaissances de niveau supérieur sur le modèle d'objets. Pour de nombreuses applications, la segmentation se résume à trouver un objet dans une image donnée. Cela implique le partitionnement de l'image en deux classes de régions uniquement. La segmentation d'image reste souvent une étape préalable et essentielle pour une analyse plus approfondie de l'image, la représentation de l'objet ou la visualisation.

9.2.1 Segmentation des images médicales

Comme la segmentation joue un rôle central dans la récupération des informations significatives à partir d'images, l'extraction efficace de toutes ces informations et des caractéristiques des images multidimensionnelles est de plus en plus importante. Dans leur forme brute, les images médicales sont représentées par des tableaux de valeurs représentant des quantités qui montrent le contraste entre les différents types de tissus du corps. Le traitement et l'analyse des images médicales sont utiles pour transformer cette information brute en une forme symbolique quantifiable. L'extraction de cette information quantitative significative peut aider au diagnostic, ainsi que dans l'intégration de données complémentaires provenant de multiples modalités d'imagerie. Par conséquent, dans l'analyse d'images médicales, la segmentation a une grande valeur clinique car elle est souvent la première étape dans l'analyse quantitative de l'image.

Néanmoins, la segmentation d'images médicales se distingue des tâches de segmentation d'images classiques et reste généralement difficile. Premièrement, de nombreuses modalités d'imagerie médicale produisent des images très buitées et floues en raison de leurs mécanismes d'imagerie intrinsèques. Deuxièmement, les images médicales peuvent être relativement mal échantillonnés. De nombreux voxels peuvent contenir plus d'un seul type de tissu, (effet de volume partiel). Dans ce cas, la perte de contraste entre deux tissus adjacents rend plus difficile leur délimitation. En plus de ces effets, certains tissus ou organes d'intérêts partagent des niveaux d'intensité similaires avec les régions voisines, conduisant à une absence de limites francches des objets. Cela implique que ces structures d'intérêt restent très difficiles à isoler de leur environnement. Par ailleurs, si l'objet à une forme complexe, ce manque de contraste sur ses limites rend la segmentation encore plus fastidieuse. Enfin, en plus des informations de l'image, une connaissance approfondie de l'anatomie et de la pathologie peut s'avérer importante pour segmenter les images médicales. L'expertise médicale est donc nécessaire afin de mieux comprendre et interpréter l'image de sorte que les algorithmes de segmentation puissent répondre aux besoins du clinicien.

Les approches initiales pour segmenter le cerveau en IRM se sont principalement concentrées sur la classification du cerveau en trois classes principales : la substance blanche (SB), la substance grise (SG) et le liquide céphalo-rachidien (LCR) [34]. Des méthodes plus récentes intégrent la segmentation des tumeurs et régions adjacentes, telles que les zones nécrotiques [37]. Ces méthodes ne sont basées que sur l'intensité du signal. Cependant, la segmentation des structures sous-corticales (à savoir les OARs) peut difficilement être réalisée uniquement sur la base de l'intensité du signal, en raison des faibles limites visibles et des valeurs d'intensité similaires entre les différentes structures sous-corticales. Par conséquent, des informations additionnelles, telles qu'un a priori de forme, de apparence ou de localisation, sont nécessaires pour effectuer la segmentation.

Parmi les techniques de segmentation qui ont exploité cette information, on peut citer : les méthodes basées sur les "atlas", les méthodes statistiques, les modèles déformables et les techniques basées sur l'apprentissage (Machine learning).

9.2.2 Methodes basées sur les atlas

Les méthodes basées sur les atlas sont largement utilisées pour l'exploitation de connaissances a priori. Un "atlas" est une image d'un patient "type" préalablement segmentée et qui sert de référence à l'image du patient à segmenter. Ces informations anatomiques sont exploitées au moyen des atlas pour être adaptés au patient en cours d'examen. La procédure générale pour effectuer des segmentations sur les images d'un patient en utilisant un ou plusieurs atlas respecte le plus souvent le même principe : recalage et propagation des contours. Tout d'abord, un champ de déformation qui met en correspondance l'atlas avec l'image du patient à segmenter est calculé en utilisant des méthodes de recalage appropriées [85]. En second lieu, le champ de déformation ainsi calculé est appliqué aux structures d'intérêt déjà segmentées sur les atlas vers l'image originale.

Presque toutes les techniques basées atlas exigent une recalage d'images durant la phase initiale. Cela signifie que le succès de la propagation des atlas dépend fortement de l'étape de recalage. L'utilisation d'un seul atlas pour propager des structures segmentées au sein d'un seul patient est généralement suffisante. Cependant, compte tenu de la grande variabilité inter-individuelle, l'utilisation d'un seul atlas peut conduire à des résultats insatisfaisants. L'utilisation de plus d'un atlas améliore la qualité de la segmentation dans ces situations. En augmentant le nombre d'atlas dans la base de données, la méthode devient plus représentative de la population et donc plus robuste lors du traitement des patients qui présentant des variations anatomiques. Cependant, lors de l'utilisation de plusieurs atlas, le point clé est de déterminer quel atlas doit être utilisé pour un patient donné. Pour ce faire, certains paramètres de similitude sont utilisés après l'étape de recalage afin de sélectionner l'atlas le plus "proche" parmi toutes les autres dans la base de données. Comme alternative à la sélection des atlas les plus proches de l'image cible, plusieurs atlas peuvent être propagés, conduisant à plusieurs solutions de segmentation, fusionnées à la fin du processus. La fusion des solutions peut finalement générer des artefacts, notamment des organes discontinus non représentatif de l'anatomie. Du point de vue clinique, la nécessité de corriger manuellement les contours automatiques a fait l'objet de plusieurs évaluations cliniques récentes [98].

Une des principales limitations des méthodes basées sur l'atlas est que les connaissances a priori incluses dans les contours du modèle ne sont pas exploitées. Pour effectuer la segmentation, ces contours sont simplement déformés. En conséquence, la plupart de l'information intégrée dans les contours, telles que la forme ou l'apparence, reste implicite et probablement sous-exploitée. Les modèles statistiques sont une alternative qui abordent cette problématique en faisant une exploitation plus explicite de ces informations pour aider à la segmentation d'images. Contrairement, aux atlas, les images ne sont pas recalées, mais les formes et, parfois, l'aspect de l'organe, sont appris afin d'être identifiés sur une image cible.

9.2.3 Modèles statistiques

Les modèles statistiques (MS) ont largement été utilisés dans le domaine de la vision par ordinateur et de la segmentation des images médicales au cours de la dernière décennie [48,58–64,99–113]. Fondamentalement, les MS utilisent une connaissance a priori de la forme par apprentissage de sa variabilité observée sur une base de données convenablement annotée. L'espace de recherche est contraint par le modèle ainsi défini. La procédure basique de MS de forme et/ou de texture est : 1) les sommets ou points de contrôle d'une structure sont modélisés comme une distribution gaussienne multivariée; 2) la forme et la texture sont modélisées en termes de moyenne et de vecteurs propres; 3) des nouvelles instances du contour sont générées grâce aux modes de variations définis par les vecteurs propres et en respectant les contraintes des sous-espaces de formes et de textures acceptables. Par conséquent, si la taille de la forme à segmenter est supérieure à la taille des données d'apprentissage, les seules formes et textures acceptables sont des combinaisons linéaires des données d'apprentissage initial.

Contrairement aux méthodes de segmentation basées sur atlas, les modèles statistiques ont besoin d'un modèle d'apprentissage. Les formes moyennes, les textures et leurs modes de variations qui définissent ce modèle sont appris à partir de d'une base de données manuellement segmentée. Si le nombre d'échantillons utilisés pour construire le modèle d'apprentissage est insuffisant, il y a un risque important de surestimation de la segmentation. De plus, la présence de bruit sur les images de la base d'apprentissage affecte la robustesse lors de la segmentation des images cibles.

Selon l'objet cible, des points du contour sont recherchés au voisinage de la forme en respectant des contraintes locales. Ainsi, une initialisation manuelle suffisamment précise doit être réalisée afin de faire converger le modèle vers la forme cible. Cette initialisation peut être fournie soit par interaction directe de l'utilisateur soit par des techniques automatiques. Cependant, si la position initiale est trop éloignée de l'objet recherché, en termes de translation, rotation ou d'échelle, cela peut conduire à une mauvaise identification d'objet.

9.2.4 Modèles Deformables

Le terme "modèle déformable" (MD) a initialement été utilisé par Terzopoulos et al. [115] pour se référer à des courbes ou surfaces, définies dans le domaine de l'image, et qui sont déformés sous l'influence de forces internes et externes. Les forces internes sont définies sur les propriétés de la courbe afin de préserver le lissage des contours pendant le processus de déformation. Les forces externes quant à elles permettent de déformer le contour en fonction des caractéristiques de l'image dans son voisinage afin de faire évoluer le modèle vers la structure d'intérêt. Par conséquent, les MD abordent le problème de la segmentation par la recherche d'une limite de l'objet vu comme une structure unique et connecté. Ces modèles peuvent être divisés en deux grandes catégories: paramétrage explicite, basés sur des représentations en maillage, et paramétrage implicite, ensembles de niveau (level-sets), représentés comme une isovaleur d'une fonction scalaire dans un espace de dimension supérieure.

Contrairement aux modèles statistiques, aucun apprentissage ou connaissance a priori n'est nécessaire pour ces modèles déformables. Ils peuvent évoluer vers la forme souhaitée, démontrant une plus grande souplesse que les autres méthodes. Néanmoins, la définition des critères d'arrêt est difficile, et elle dépend des caractéristiques du problème. Les modèles déformables paramétriques ont été utilisés avec succès dans un large éventail d'applications et de problèmes. Une propriété importante de ce type de représentation est sa capacité à représenter les limites à une résolution infra-pixel, ce qui est essentiel dans la segmentation des structures minces. Cependant, ils présentent deux limitations principales. Tout d'abord, si la variation de la taille et de la forme entre le modèle initial et l'objet cible est importante, le modèle doit être paramétré dynamiquement pour récupérer fidèlement la limite de l'objet. La seconde limitation est liée aux complications qu'elles présentent pour faire face aux changements topologiques, telles que le fractionnement ou la fusion de parties du modèle. Les modèles géométriques fournissent une solution élégante pour répondre à ces limitations car, en se basant sur théorie de l'évolution de la courbe, courbes et surfaces évoluent indépendamment du paramétrage. Cela permet une gestion automatique des transitions topologiques.

Un inconvénient commun aux deux modèles, géométriques et paramétriques, est que les images auxquelles appliquer l'un de ces modèles doivent avoir des bords suffisamment nets et des régions homogènes pour une modélisation explicite. En conséquence, les modèles déformables traditionnels ne parviennent généralement pas segmenter en présence d'inhomogénéités d'intensité importantes et/ou de faibles contrastes.

9.2.5 Machine Learning

L'apprentissage automatique ou apprentissage statistique (machine learning en anglais) a été largement utilisé dans le domaine de l'analyse IRM presque depuis sa création. Ces méthodes de segmentation bases sur un apprentissage supervisé d'abord extraient caractéristiques de l'image avec des informations souvent plus riches que des information de nievau de gris seule. Puis ils construisent un modèle de classification basé sur les caractéristiques de l'image en utilisant des algorithmes d'apprentissage supervisé.

Parmi toutes les informations possibles qui peuvent être extraites afin de segmenter des structures cérébrales dans les images médicales, les plus couramment utilisées sont : basées sur le niveau de gris, basées sur la probabilité et l'information spatiale. Elles représentent les cas les plus simples de caractéristiques. Les caractéristiques basées sur le niveau de gris exploitent le niveau de gris d'un voxel et l'aspect de son voisinage. Dans sa représentation la plus simple, des patches carrés autour d'un pixel ou d'un voxel sont utilisées en 2D et 3D, respectivement, avec des valeurs de taille de patch typique allant de 3 à 9 pixels ou voxels. Les caractéristiques basées sur la probabilité analysent la probabilité d'un voxel d'appartenir à une structure déterminée. La carte qui contient ces probabilités est créée à partir d'une base de données préalablement annotée. En plus du niveau de gris et de la probabilité, la localisation du voxel dans l'espace de l'image peut également être utilisée.

L'objectif de nombreux algorithmes d'apprentissage consiste à rechercher une famille de fonctions afin d'identifier un membre de la famille mentionnée qui minimise un critère d'apprentissage. Les réseaux de neurones artificiels (en anglais Artificial Neural Network, ANN) et les machines à vecteurs de support ou séparateurs à vaste marge (en anglais Support Vector Machine, SVM) sont parmi les méthodes d'apprentissage les plus populaires utilisées non seulement pour la segmentation des structures anatomiques du cerveau [42, 43, 47, 73–76, 126–128], mais aussi pour la classification des tumeurs [129–131] ou le diagnostic automatique [132].

ANN représente un système de traitement d'informations comportant un grand nombre de composants interconnectés de traitement individuels, à savoir les neurones. Motivé par la façon dont le cerveau humain traite les informations d'entrée, les neurones travaillent ensemble d'une manière distribuée à l'intérieur de chaque réseau pour apprendre des connaissances d'entrée, traiter ces informations et de générer une réponse significative. Chaque neurone n dans le réseau traite l'entrée grâce à l'utilisation de son propre poids w_n, une valeur biais b_n, et une fonction de transfert qui prend la somme de w_n et b_n. En raison de leur efficacité dans la résolution de problèmes d'optimisation, ANNs ont été largement intégrés dans les algorithmes de segmentation pour définir les structures sous-corticales [42, 73, 74, 76, 126, 128]..

Fondamentalement, l'idée principale derrière SVM est de trouver le plus grand hyperplan de marge qui sépare deux classes. La distance minimale de l'hyperplan de séparation entre deux classes est appelée marge. Ainsi, l'hyperplan optimal est celui qui fournit la marge maximale, représentant la plus grande séparation entre les classes. En transformants les objets de leur espace d'origine vers un espace de caractéristiques de dimension supérieure [143], SVM peut séparer les objets qui ne sont pas linéairement séparables. Leur bonne capacité de généralisation et leur capacité à classer correctement les données non-linéairement séparables ont conduit à un intérêt croissant sur eux pour les problèmes de classification.

En introduisant des méthodes de Machine Learning, les algorithmes développés pour le traitement d'images médicales deviennent souvent plus "intelligents" que les techniques conventionnelles. Les techniques Machine Learning ont montré de meilleures performances que les autres approches plus traditionnelles pour la segmentation segmentant des structures cérébrales [43, 47, 75, 76]. Les développements récents des techniques d'acquisition d'imagerie médicale ont conduit à une augmentation de la complexité de l'analyse des images. Cela apporte de nouveaux défis où l'analyse manuelle d'une grande quantité de données est limitée. Dans ce contexte, les techniques Machine Learning nous semblent les plus adaptées pour faire face à ces nouveaux défis. Par ailleurs, un nouveau domaine de l'apprentissage automatique a récemment émergé avec l'intention de rapprocher le Machine Learning de ses objectifs initiaux : l'intelligence artificielle. Il s'agit du Deep Learning. Les progrès récents sur l'utilisation des réseaux profonds pour la reconnaissance d'image, reconnaissance de la parole, ou d'autres applications ont montré qu'ils offrent actuellement les meilleures solutions à bon nombre de ces problèmes. Par conséquent, nous allons considérer l'utilisation du Deep Learning pour résoudre le problème de la segmentation des structures cérébrales en radiothérapie.

9.3 Contribution

9.3.1 Deep Learning

Le Deep learning est un nouveau sous-domaine du machine learning qui met l'accent sur l'apprentissage des modèles hiérarchiques de données. L'étude du deep learning moderne prend beaucoup de son inspiration dans la recherche des ANN des décennies précédentes. La plupart des algorithmes d'apprentissage actuels correspondent aux architectures peu profondes avec 1 jusqu'à 3 niveaux d'abstraction. Inspirés par l'architecture "profonde" du cerveau, les chercheurs dans les domaines des réseaux de neurones ont tenté pendant des décennies de former des ANN multicouches profonds. Néanmoins, les premières tentatives rencontrant un succès n'ont été publiées qu'à partir de 2006. Malgré les résultats remarquables des ANN pour effectuer certaines tâches [147], d'autres approches ont dominé pendant les années 90 et 2000 [141, 143, 148]. L'une des principales raisons de l'abandon des ANN en faveur de ces approches est la difficulté de former des réseaux profonds. L'apprentissage d'architectures profondes est une tâche difficile et les méthodes classiques qui ont prouvé leur efficacité lors d'application à des architectures peu profondes ne sont plus adaptées. Finalement, le simple fait d'ajouter des couches ne conduit pas nécessairement à de meilleures solutions. Au contraire, lorsque le nombre de couches cachées augmente il devient plus difficile d'obtenir une bonne généralisation.

Par conséquent, jusqu'à récemment, la plupart des techniques de Machine Learning ont exploité des architectures peu profondes, où les réseaux étaient généralement limités à une ou deux couches cachées.

Cependant, en 2006, le concept de Greedy Layer-Wise Learning a été introduit [149, 151, 152]. Ce nouveau concept bénéficie d'une procédure d'apprentissage semi-supervisé. L'apprentissage non supervisé est utilisé dans une première étape pour initialiser les paramètres des couches, une couche à la fois, et puis un réglage fin de l'ensemble du système se fait par une tâche supervisée. Depuis, le deep learning a émergé comme un nouveau domaine de recherche du Machine Learning, avec un fort impact sur un large éventail de domaines de recherche [151,153].

L'un des avantages du deep learning par rapport aux ANN peu profonds est que des fonctions complexes peuvent souvent être estimées avec la même précision en utilisant un réseau plus profond mais avec beaucoup moins d'unités par rapport à un réseau typique de deux "grandes" couches cachées. En outre, avec de plus petits degrés de liberté, le deep learning nécessite des ensembles de données plus petits pour l'apprentissage. Un autre facteur, probablement plus convaincant, est que les approches typiques de classification doivent être généralement précédés par une étape de sélection de caractéristiques, où les caractéristiques les plus discriminantes sont privilégiées pour un problème donné. Les approches de deep learning quant à elles, ont la capacité d'apprendre automatiquement les caractéristiques des données. Cette spécificité a largement contribué à l'amélioration en termes de précision.

Parmi les différentes techniques de deep learning disponibles, nous utiliserons Auto-encoders (AE). Dans sa représentation la plus simple, un AE se compose de deux éléments: un codeur $\mathrm{h}(\cdot)$ et un decodeur $\mathrm{g}(\cdot)$. Tandis que le codeur transforme l'entrée à une certaine représentation cachée, le décodeur transforme la représentation cachée à une version reconstruite de l'entrée x. Un AE est donc formé pour minimiser la contradiction entre les données et sa reconstruction. Néanmoins, si aucune autre restriction outre la minimisation d'erreur n'est imposée, l'AE peut potentiellement n'apprendre que la fonction identité. Une solution pour éviter cela est d'ajouter un processus aléatoire dans la transformation de l'entrée à sa reconstruction, il s'agit du Denoising Auto-encodeurs (DAE) [161–166].

En général, un DAE est implémenté comme un réseau neuronal d'une couche cachée formée pour reconstruire un point x à partir de sa version corrompue x. Par conséquent, un AE est converti en un DAE, en ajoutant simplement une étape de corruption stochastique modifiant l'entrée. Par exemple, dans [161], le processus de corruption stochastique consiste à mettre au hasard quelques-unes des entrées à zéro. Plusieurs DAE peuvent être empilés pour former un réseau profond en alimentant la représentation cachée d'un DAE de la couche inférieure alimentant lui même l'entrée de la couche suivante [159]. Il s'agit alors de Stacked Denoising Auto Encoder (SDAE).

L'apprentissage du SDAE est composé de deux étapes : apprentissage non-supervisé et supervisé. Les poids entre les couches du réseau sont d'abord appris par l'étape de pré-apprentissage non supervisé. Le pré-apprentisage non-supervisé de l'architecture proposée est réalisé une couche à la fois. Chaque couche est formée en tant que DAE, en minimisant l'erreur de reconstruction de son entrée. Le DAE de la couche supérieur utilise alors la sortie du DAE de niveau inférieur comme entrée. Une fois que les premières couches k sont formées, la couche $k{+}1$ peut être formée parce que la représentation latente de la couche inférieure peut être alors calculée. Une fois que tous les poids du réseau sont calculés le réseau passe par une deuxième étape d'apprentissage appelée supervisé appelée fine-tunning, où l'erreur de prédiction est réduite sur une tâche supervisée

9.3.2 Caractéristiques utilisées pour la segmentation

Quelle que soit l'efficacité de la stratégie d'apprentissage automatique appliqué, le choix de caractéristiques pertinentes est crucial pour des problèmes de classification. Les recherches récentes sur la segmentation des structures du cerveau par des techniques de Learning Machine ont tendance à se concentrer sur l'utilisation de plusieurs algorithmes d'apprentissage plutôt que dans l'ajout de caractéristiques plus discriminantes dans le système. Les caractéristiques traditionnelles, introduites précédemment, ont souvent été utilisées lors de la segmentation de structures cérébrales avec un succès considérable. Cependant, l'utilisation des caractéristiques alternatives peut : i) améliorer les performances de classification, ii) réduire, dans certains cas, le nombre de caractéristiques utilisées pour décrire les informations de texture d'une région donnée. En dehors de l'application de SDAE au problème de la segmentation des OARs, l'une des principales contributions de ce travail est l'utilisation des caractéristiques qui ne sont pas encore utilisées pour la segmentation de ces structures du cerveau.

Parmi l'ensemble des OARs impliqués dans la RT, certains présentent une homogénéité de texture plus importante et une variation plus limitée de la forme que d'autres. Dans ce premier groupe, nous pouvons inclure le tronc cérébral, les yeux et le cristallin. A l'inverse, d'autres OARs ont une texture plus hétérogène des variations inter-individuelles plus importantes en termes de taille ou de localisation. Ce deuxième groupe est constitué par les nerfs optiques, le chiasma, l'hypophyse et la tige pituitaire. En raison des différences entre les caractéristiques des deux groupes, certaines des caractéristiques proposées dépendent de l'organe à segmenter et ne sont pas adaptées à tous les organes étudiés dans ce travail. Alors que la segmentation de certains organes exploite l'utilisation d'une carte de distances géodésiques et de descripteurs fondés sur les motifs binaires locaux en 3D pour obtenir de meilleurs résultats (groupe A), la segmentation d'autres OARs utilise la texture et l'analyse contextuelle (groupe B). Le gradient d'image est exploré dans ce groupe. Parmi toutes les caractéristiques qui peuvent être extraites de l'analyse de texture, nous utilisons les suivantes : la moyenne, la variance, l'asymétrie, l'aplatissement, l'énergie et l'entropie. De plus, la décomposition discrète en ondelettes compos également le vecteur de caractéristiques.

9.3.3 L'apprentissage

Un pré-traitement est appliqué extraire les caractéristiques à l'ensemble des patients. Toutes les images sont redimensionnées à la résolution 1 x 1 x 1 mm^3. Toutes les images IRM T1 sont spatialement alignées de telle sorte que la ligne de la commissure antérieure et la commissure postérieure (AC-PC) est orientée horizontalement dans le plan sagittal, et la fissure inter hémisphérique est alignée sur les deux autres axes. Ce procédé représente donc également l'étape d'initialisation pour la segmentation d'un nouveau patient. Enfin, les images sont normalisées.

Une carte de probabilité et un masque de recherche sont créés lors de la phase d'apprentissage. Les zones d'intérêts manuellement contourées sur l'ensemble des données d'apprentissage sont sommées dans un volume pour créer une carte de probabilité pour chaque OAR. Cette carte contient ainsi des valeurs continues par voxel dans l'intervalle [0,1], indiquant la fréquence à laquelle un organe apparaît dans l'ensemble de données. Cette valeur indique la probabilité qu'un voxel donné appartienne à une structure. La carte de probabilité est également utilisée pour réduire le nombre d'échantillons qui sont introduits dans le classificateur. De cette carte, une région d'intérêt (ROI) est générée. Le critère d'élagage est basée sur la probabilité d'un voxel d'appartenir à une quelconque des structures d'intérêt. Ainsi, tout voxel contenant une probabilité supérieure à zéro est pris en compte pour créer le masque de recherche, pour chaque structure, et qui sera utilisé pour élaguer les voxels dans l'étape d'extraction de caractéristiques. Pour être sur de que les OARs des nouveaux patients seront à l'intérieur de ce masque commun une marge de sécurité est générée par l'application d'une dilatation morphologique.

9.3.4 Classification

La classification est faite à une classe à la fois. Cela signifie qu'un classifieur binaire est utilisé pour chacune des structures. Le pré-traitement de l'image à segmenter est le même que celui utilisé pendant l'apprentissage. Pour l'extraction des caractéristiques, la carte de probabilités et la ROI de recherche crées pendant l'apprentissage sont aussi utilisées. Les valeurs des caractéristiques pour la classification sont mises à l'échelle en concordance avec la mise à l'échelle des valeurs utilisées lors de la phase de d'apprentissage.

9.4 Matériels et méthodes

Tout le code qui a été utilisé dans cette thèse a été mis en œuvre en utilisant les plates-formes suivantes: MATLAB (The MathWorks Inc., Natick, MA, 2000) et Microsoft Visual Studio (MSVC) 2010.

Les examens IRM de 15 patients pris en charge dans le cadre d'une radiochirurgie Leksell Gamma Knife ont été utilisés dans ce travail. Au total, quatre experts ont participé à aux sessions de contourage manuel des OARs. Ces contours manuels ont été utilisés pour créer les contours de référence. Les contours de référence ont été obtenus dans cette thèse en utilisant le concept de calcul de cartes de probabilité. Les cartes de probabilité sont seuillées à un niveau variable afin de créer le masque. Le seuil a été fixé à 50% ou à 75%, en fonction du nombre d'experts impliqués (3 ou 4 selon l'OAR étudié).

Les techniques typiques de validation pour évaluer la performance d'un classifieur s'appuient sur un partage des données en deux groupes : apprentissage et test. Compte-tenu du nombre limité d'examens dans cette thèse, nous avons utilisé la méthode Leave-one-out Cross validation (LOOCV). Cette technique consiste à utiliser un seul patients pour le test et les autres pour l'apprentissage. Ce processus est répété autant de fois que de patients disponibles, à savoir 15. Ainsi, à chaque itération, 14 examens sont utilisés pour l'apprentisage et 1 pour la classification.

Plusieurs critères permettent d'évaluer la qualité de segmentation d'une image et selon lesquels différentes métriques d'erreur sont définies. Ces différents critères d'évaluation de segmentation ont été classés par [186] : exactitude, précision, répétabilité et efficacité. Les métriques sont estimées à partir du contour et la taille/le volume de l'objet segmenté. Chaque métrique reporte une information différente et doit être considérée dans un contexte approprié. Bien que les mesures basées sur le volume, telles que Dice Similarity Coefficient (DSC) sont largement utilisées pour comparer les similitudes entre volumes, elles sont assez peu sensibles aux différences sur les bords lorsque ces différences ont un faible impact sur le volume global. Ainsi, les mesures fondées sur la distance, telles que la distance de Hausdorff, sont également utilisées pour évaluer la qualité d'une segmentation. En complément, la sensibilité et la spécificité sont aussi évaluées. Finalement, pour évaluer l'efficacité, le temps nécessaire à l'exécution de l'algorithme est également mesuré et analysé. A titre de comparaison à des algorithmes de référence dans la littérature, un autre classifieur basé sur SVM a été étudié. En outre, les segmentations automatiques générées par notre système de deep learning, sont comparées aux segmentations manuelles obtenues par les experts. A partir des valeurs des métriques ainsi recueillies une analyse statistique est réalisée afin de démontrer que les différences de volume et de surface étaient significativement différentes entre le system basé sur SVM et notre système de classification basé sur SDAE. D'autre part, nous avons également réalisé une analyse statistique entre le résultat des segmentations manuelles et les contours générés par notre système. Dans ce cas, nous souhaitons prouver que, bien que les résultats de certains observateurs manuels ont été meilleurs que les résultats fournis par notre approche, les différences ne sont pas significatives.

9.5 Résultats et discussion

Selon les caractéristiques intrinsèques de certains OARs, les résultats ont été séparés en deux groupes. Les résultats du système proposé ont été comparés avec une approche de Learning Machine largement utilisé pour la classification : SVM. En plus des caractéristiques usuelles basées sur l'espace et l'intensité, nous avons ajouté des nouvelles caractéristiques. Ces caractéristiques sont généralement dépendant des organes, et leur évaluation a ainsi été réalisée en conséquence selon deux groupes A et B.

Les résultats fournis dans ce travail montrent que le système de classification proposé surpasse toutes les configurations des classifieurs SVM ou SDAE avec les caractéristiques usuelles. L'ajout de nouvelles caractéristiques dans chaque groupe a augmenté la similitude entre volumes tout en réduisant la distance de Hausdorff. Pour tous les OARs, les schémas de classification proposées pour les groupes A et B ont obtenu les meilleurs résultats pour les différences mesures de similarité, de surface et de volume. La sensibilité et la spécificité sont également améliorés par l'utilisation du système de classification proposé. Premièrement, les valeurs de sensibilité étaient plus élevées dans les configurations basées sur SDAE que dans configurations basées sur SVM. Deuxièmement, l'inclusion de nouvelles caractéristiques dans le système de classification a améliorée des valeurs de sensibilité par rapport aux autres configurations dans les systèmes basées sur SDAE. Cette tendance a été identifiée pour tous les OARs des deux groupes. Les valeurs de spécificité obtenus par les systèmes proposés dans les deux groupes étaient dans environ la moitié des cas parmi ceux les mieux classés.

L'analyse statistique sur les segmentations automatiques a démontré que les résultats obtenus par le système proposé sont significativement meilleurs que les autres groupes.

En ce qui concerne la comparaison aux annotations manuelles, l'erreur de segmentation que nous avons obtenu est comparable à celle obtenue entre les observateurs lorsque les contours sont délimités sans contraintes de temps. Dans ces comparaisons, nous pouvons observer que les résultats de segmentation générés par l'approche proposée sont distribues avec la même variabilité que les experts dans la plupart des cas. L'analyse statistique sur ces résultats du système de classification comparés aux contours manuels souligne que les différences ne sont généralement pas statistiquement significative. En outre, dans certains cas où les différences sont significatives, notre classificateur automatique offre un meilleur résultat que le contourage manuel. Nous pouvons ainsi conclure que les contours automatiques générés par le système de classification proposé sont similaires aux annotations manuelles.

Les résultats démontrent également que le système de deep learning proposé reposant sur l'apprentissage a surpassé tous les travaux précédents lorsque l'on s'intéresse à l'ensemble des OARs analysés. Bien qu'il n'a pas été possible dans ce travail d'utiliser les mêmes ensembles de données que celles utilisées dans les études précédentes, les performances plus élevées de notre approche, comme indiqué par les résultats, suggère sa supériorité pour segmenter ces structures. Les résultats montrent qu'en utilisant SDAE comme classificateur, le temps de segmentation a été significativement réduit par rapport à d'autres méthodes classiques de machine learning, telles que SVM. Cela est particulièrement remarquable si on tient en compte le fait que la plupart des travaux en références dans cette thèse pour segmenter les OARs sont basés sur des techniques atlas, et donc dépendantes du recalage. Cette étape de recalage rend la segmentation chronophage en comparaison de l'approche proposée.

L'implémentation actuelle du système proposé n'est pas optimisée informatiquement et goulot d'étranglement du processus reste l'étape d'extraction de caractéristiques. Cependant son temps de traitement varie entre 1 à 6 secondes pour chacun des OARs. Bien qu'elle ne soit pas une étape très coûteuse en temps de calcul, elle représente plus de 95% du temps total de la segmentation. Comme l'extraction des caractéristiques ne nécessite pas d'opérations complexes de programmation, sa parallélisation est facilement abordable. Cela peut réduire sensiblement l'ensemble du processus de segmentation jusqu'à moins d'une seconde pour un organe entier.

L'un des points forts des méthodes de deep learning repose sur leur capacité à transférer les connaissances de l'homme à la machine. Ces machines "apprennent" à partir d'un ensemble de données. Ainsi, par exemple, en l'absence de limites visibles, le classifieur utilise l'expertise des médecins transférée au système pour la réalisation de cette tâche.

Appendix A
Artificial Neural Networks

A.0.1 Artifical Neural Networks

Artificial Neural Networks (ANN) based methods provide a robust approach for approximating real-valued, discrete-valued or vector-valued targeted functions. They are massively parallel computing systems consisting of an extremely large number of simple processors with many interconnections between them. For certain types of problems, ANN are among the most effective learning methods employed during the last decades [REF?].

A.0.1.1 Biological motivation

The observation that biological learning systems are built of very complex networks of interconnected neurons inspired the study and development of ANN based systems. Thus, in a fuzzy analogy, ANN are composed by a densely interconnected set of simple units. Each of these units takes a number of real-valued inputs and produces a single real-valued output. Inputs and outputs at each neuron may represent the outputs and inputs of other units, respectively.

To develop the basis of this similarity, let us consider some certainties from neurobiology. A neuron is a special biological cell that has the ability to process information. It is estimated that the human brain contains a densely interconnected network of nearly 10^{11} of neurons. Each neuron is connected to 10^3 to 10^4 other neurons, on average. Neuron activity is commonly inhibited or excited through connections to other neurons. Neurons communicate through a very short train of pulses, typically with a duration of milliseconds. Although the fastest neuron switching times are estimated to be on the order of 10^{-3} seconds, this time is much slower than computer switching times, which are on the order of 10^{-10} seconds. However, complex decisions performed by humans can be done surprisingly quick. For instance, visually recognizing a familiar person, such as your mother or father, it requires approximately 10^{-1} seconds. Taking into account biological switching times, this implies that the sequence of neurons being excited during this 10^{-1} seconds interval cannot be longer than a few hundred of serial stages. This observation led to many researchers during the beginning of ANN to speculate that the information processing abilities of biological neural systems must follow from highly parallel processes operating on representations that were distributed over many neurons. Thus, one motivation of ANN-based systems is to capture this type of highly parallel computation based on distributed representations.

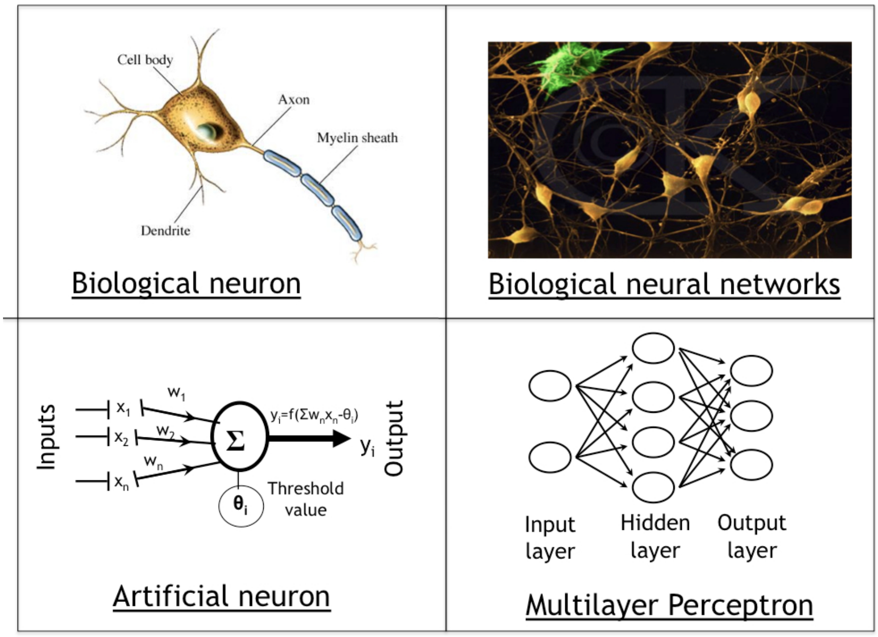
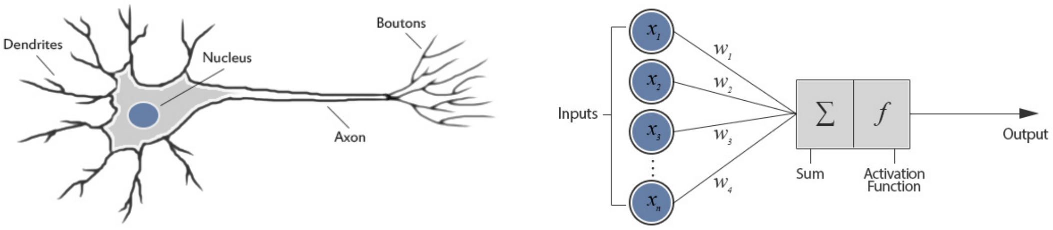


Figure A.1: Comparative schemes of biological and artificial neural system.

A.0.1.2 The basics of artificial neural networks

ANN are therefore a biologically inspired computational framework where a number of simple computational units, referred as to neurons, are connected together to compute a more complex function. The complexity of biological neurons is highly abstracted when modeling artificial neurons. These basically consist of inputs, which are multiplied by weights, and then computed by a mathematical function which determines the activation of the neuron. The activation function associated with each neuron determines how that neuron's value (or activation) is updated. A typical example of an artificial neuron is shown in figure A.2,b.

Benefit of artificial neuron model simplicity can be seen in its mathematical description defined in equations A.1 and A.2, which represent the neuron pre-activation or input activation and the neuron activation or output activation, respectively:



(a) Biological neuron

(b) Artificial neuron

Figure A.2: Appearance of a biological(a) and an artificial(b) neuron.

$$a(x) = \sum_{i=0}^{m} x_i \cdot w_i + b \tag{A.1}$$

$$y = f(\sum_{i=0}^{m} x_i \cdot w_i + b) \tag{A.2}$$

where x_i represents the i unit inputs, w_i are the weight values for each input i, b is the bias term and f is the transfer or activation function. Lastly, y represents the output value of the neuron. As seen from the artificial neuron model and its equation (A.2), the major unknown variable is its transfer function. Figure A.3 shows some of the most common activation functions employed in ANN. In each case, the x-axis represents the value of the net input whilst the y-axis is the output from the neuron. Among these activation function types, sigmoid functions are widely employed in ANN due to its remarkable computational and mathematical properties. Additionally, most biological neurons are sigmoid units, in the sense that their frequency response on input has a region of maximum sensitivity somewhere between a threshold and a saturation point. Mathematical formulation of the sigmoid activation function is described below:

$$f(a) = sigm(a) = \frac{1}{1 + exp(-a)} \tag{A.3}$$

where a denotes the pre-activation function defined in equation A.1.

Since a neural network is built out of interconnected neurons, the function of an entire neural network is simply the computation of the output of all the neurons. Training the network involves presenting the network with some sample data and modifying weights to better approximate an activation function to obtain the desired output. Even though a precise definition of learning is ambitious to formulate, a learning process in an ANN context can be viewed as the issue of updating the artificial network architecture and connection weights so that a network can efficiently perform a specific task.

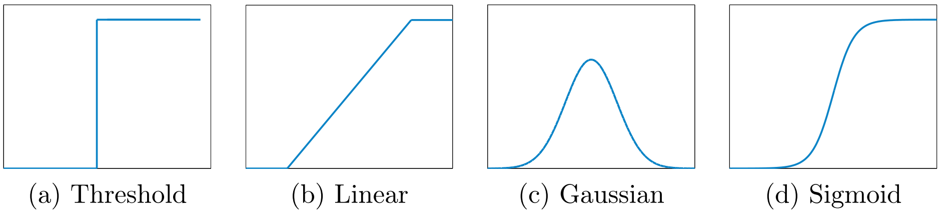

(a) Threshold (b) Linear (c) Gaussian (d) Sigmoid

Figure A.3: Common activation functions.

A.0.1.3 Multilayer feed-forward networks

A single neuron, however, is not very useful due to its limited mapping ability. Regardless of which activation function is used, the neuron is only able to represent an oriented ridge-like function, being able to only handle linearly separable or linearly independent problems. Further extensions of single neuron based networks concern models in which many neurons are interconnected and organized into layers, building blocks of a larger, much more practical structures. Neurons in the same layer are fully connected to the neurons in the previous layer, except for the first layer, because this layer is not formed by neurons but by the vector $\mathrm{x}^{(i)}$ that will be the input to the network.

Neural networks can be built following multiple and diverse architectures. According to the direction of connections between layers ANN can be grouped into two major categories: (i) feed-forward networks, in which no loops exist in the graph, and (ii) feedback networks, also known as recurrent, where loops are present due to feedback connections. Different network architecture leads to different learning algorithms. The most common choice is a n_l-layered network, where the first layer represents the input layer, layer n_l is the output layer, and each layer l is densely connected to layer $l+1$. We will discuss the former, since no other network topology will be analyzed in this dissertation.

Multilayer feed-forward (MLF) neural network represents one of the most popular multilayer ANN. In a feed forward neural network, neurons are only connected forward. Each layer of the neural network contains connections to the next layer, but there are no connections back. This means the signal flow is from input to output units, strictly in a feed-forward direction. Typically, the network consists of a set of sensory units that constitute the input layer, one or more hidden layers of computation nodes, and an output layer of computation nodes. In its common use, most neural networks will have one hidden layer, and it's very uncommon for a neural network to have more than two hidden layers. The input signal propagates through the network in a forward direction, on a layer by layer basis. In this context, to compute the output of the network, activations in primary layers are computed first, up to reach the last layer, L_{n_l}. In figure A.4 a simple MLF with 3 inputs, 1 output, and 1 hidden layer containing 3 neuron units is shown.

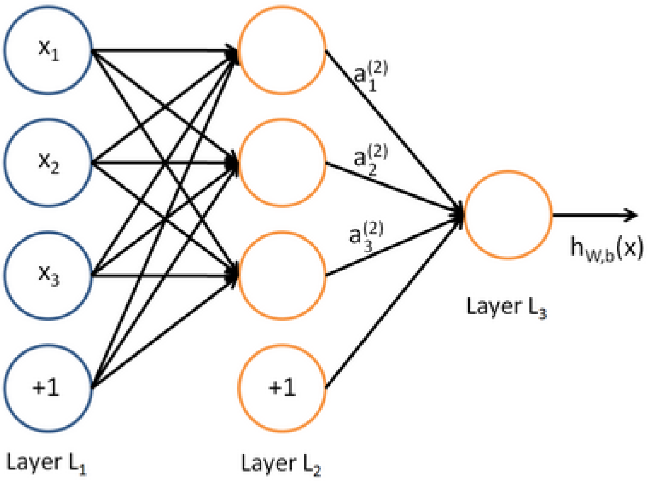


Figure A.4: A simple MLP with 3 inputs, 1 output, and 1 hidden layer containing 3 hidden units.

A.0.1.4 Training a Multilayer Network

The way in which this model learns to predict the label y$^{(i)}$ associated to the input x$^{(i)}$ is by calculating the function

$$h_{W,b}(x) = a^{(n)} = f(z^{(n)}) \tag{A.4}$$

where n is the number of layers, b is a matrix formed by $n-1$ vectors storing the bias term for the s neurons in each layer, and W is a vector of $n-1$ matrices each of which is formed by s vectors, each one representing the weight of one of the neurons in one of the layers. To achieve the learning process the training set is fed into the function in equation A.4. Calculating the value of $h_{W,b}(x)$ is called a *feedforward* pass. Therefore, to train the network, the first thing to do is to initialize the weights W and the bias term b. This should be done using random values near zero. Otherwise, all the neurons could end up firing the same activations and not converging to the solution.

Let consider the network shown in figure A.4 as example. In this setting, the number of layers, n_l is equal to 3. Each layer l is denoted as L_l, so input layer is represented by L_1, and L_3 is the output layer in our network. Let denote W_{ij}^l to refer to the weight associated with the connection between the unit j in layer l, and the unit i in layer $l+1$. In addition, b_i^l is used to represent the bias associated with the unit i in layer $l+1$. The output value, also known as activation, of a unit i in layer l is denoted by a_i^l. Therefore, for the first layer, the activation, a_i^1 is simply the i-th input. Thus, given a fixed setting of the parameters W and b, the neural network defines a hypothesis $h_{W,b}(x)$ (equation A.4), which output is a real number. Particularly, computation of our neural network is represented by:

$$a_1^{(2)} = f(W_{11}^{(1)}x_1 + W_{12}^{(1)}x_2 + W_{13}^{(1)}x_3 + b_1^{(1)}) \tag{A.5}$$

$$a_2^{(2)} = f(W_{21}^{(1)}x_1 + W_{22}^{(1)}x_2 + W_{23}^{(1)}x_3 + b_2^{(1)}) \tag{A.6}$$

$$a_3^{(2)} = f(W_{31}^{(1)}x_1 + W_{32}^{(1)}x_2 + W_{33}^{(1)}x_3 + b_3^{(1)}) \tag{A.7}$$

$$h_{W,b}(x) = a_1^{(3)} = f(W_{11}^{(2)}a_1^{(2)} + W_{12}^{(2)}a_2^{(2)} + W_{13}^{(2)}a_3^{(2)} + b_1^{(2)}) \tag{A.8}$$

Thus, we have in this network that $W^{(1)} \in \mathbb{R}^{3x3}$ and $W^{(2)} \in \mathbb{R}^{1x3}$.

If we now let $z_i^{(l)}$ denote the total weighted sum of inputs to unit i in layer l, including the bias term:

$$z_i^{(2)} = \sum_{j=1}^{n} W_{ij}^{(1)}x_j + b_i^{(1)} \tag{A.9}$$

activation of unit i in layer j can be reformulated in a more compact notation as:

$$a_i^{(l)} = f(z_i^l) \tag{A.10}$$

Extension of the activation function $f(\cdot)$ to be applied to vectors in an element-wise function (i.e., $f([z_1, z_2, z_3]) = [f(z_1, f(z_2, f(z_3)])$ will allow equations (A.5-A.8) to be reformulated as:

$$z^{(2)} = W^{(1)}x + b^{(1)} \tag{A.11}$$

$$a^{(2)} = f(z^{(2)}) \tag{A.12}$$

$$z^{(3)} = W^{(2)}a^{(2)} + b^{(2)} \tag{A.13}$$

$$h_{W,b}(x) = a^{(3)} = f(z^{(3)}) \tag{A.14}$$

More generally, in order to calculate activations of layer $l+1$, $a^{(l+1)}$, we need to calculate, for each layer l, starting with $l = 1$ and knowing that $a^{(1)} = x$

$$z^{(l+1)} = W^{(l)}a^{(l)} + b^{(l)}, \tag{A.15}$$

$$a^{(l+1)} = f(z^{(l+1)}) \tag{A.16}$$

Once we have produced a *feed-forward* pass, we need to calculate the cost function. We define the cost function of a single training example (x, y) as

$$J(W, b; x, y) = \frac{1}{2}\|y - h_{W,b}(x)\|^2 \tag{A.17}$$

that is, half of the squared distance from the prediction to the ground truth. For a whole training set $(x^{(1)}, y^{(1)}), ..., (x^{(m)}, y^{(m)})$ we will use

$$J(W, b) = \frac{1}{m}\sum_{i=1}^{m} J(W, b; x^{(i)}, y^{(i)}) + \frac{\lambda}{2}\sum_{l=1}^{n-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ij}^{(l)})^2 \tag{A.18}$$

where m is the number of examples and λ is the weight decay parameter. This parameter λ helps to prevent overfitting by penalizing the cost when the weights grow too much.

Now that we have a function that measures the cost of all predictions with a particular set of weights, we need a way to update those weights so that, in next iteration, the cost will be reduced and the training may converge to a minimum, hopefully the global one. This update value is:

$$\begin{aligned} \nabla W = \frac{\partial}{\partial W^{(l)}} J(W, b) = [\frac{1}{m}\sum_{i=1}^{m} \nabla_{W^{(l)}} J(W, b; x^{(i)}, y^{(i)})] + \lambda W^{(l)} \\ [\frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial W^{(l)}} J(W, b; x^{(i)}, y^{(i)})] + \lambda W^{(l)} \end{aligned} \tag{A.19}$$

$$\begin{aligned} \nabla b = \frac{\partial}{\partial b^{(l)}} J(W, b) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{b^{(l)}} J(W, b; x^{(i)}, y^{(i)})] \\ \frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial b^{(l)}} J(W, b; x^{(i)}, y^{(i)}) \end{aligned} \tag{A.20}$$

Therefore, the first step is to calculate $\nabla_{W^{(l)}} J(W, b; x^{(i)}, y^{(i)})$ and $\nabla_{b^{(l)}} J(W, b; x^{(i)}, y^{(i)})$ for each example independently. This step is done with the backpropagation algorithm.

A.0.1.5 Backpropagation Algorithm

Backpropagation is a method of supervised learning often used to train feed-forward neural networks. Its use allows to calculate the factors by which each weight should be updated in order to minimize the error produced between the prediction and the ground truth given a set of weights W and a bias term b. It proceeds as follows:

- Perform a feed-forward pass, that is, calculate the final activations $h_{W,b}(x) = a^{(n)}$, where n is the number of layers, and denoting that $a^{(n)}$ are the activations of the last layer. This will give us a vector of predictions achieved by the actual weights θ. Moreover, store all the intermediate $z^{(l)}$ and $a^{(l)}$ for each layer l for a later use.
- For each final activation $a_i^{(n)}$ with $i = 1, ..., l$, calculate the penalization term

$$\delta_i^{(n)} = \frac{\partial}{\partial z^{(n)}} \frac{1}{2} \|y - h_{W,b}(x)\|^2 = -(y_i - a_i^{(n)}) \cdot f'(z_i^{(n)}) \tag{A.21}$$

  This factor indicates how different the prediction of the model is from the ground truth.

- Propagate the penalization term to the previous layers by calculating for each node i in layer l except the first layer, because the input does not need to be corrected

$$\delta_i^{(l)} = ((W_i^{(l)})^T \delta_i^{(l+1)}) \cdot f'(z_i^{(l)}) \tag{A.22}$$

- Finally, compute the partial derivatives

$$\nabla_{W^{(l)}} J(W, b; x, y) = \delta_i^{(l+1)} (a^{(l)})^T \tag{A.23}$$

$$\nabla_{b^{(l)}} J(W, b; x, y) = \delta_i^{(l+1)} \tag{A.24}$$

Now, we can calculate ∇W and ∇b with the formulas in the previous section (equations A.19 and A.20,respectively). These partial derivatives should now be used to properly update the old weights with some optimization technique such as gradient descent, conjugate gradient or L-BFGS algorithm [194], for example.

Appendix B
Support Vector Machines

B.0.2 Support Vector Machines

Another widely employed ML system, which also represents a state-of-the-art classifier, is Support Vector Machines (SVM). It was originally proposed by Vapnik [141] and [142] for binary classification. In contrast with other machine learning approaches like artificial neural network which aims at reducing empirical risk, SVM implements the structural risk minimization (SRM) that minimizes the upper bound of generation error.

Support vector machines and their variants and extensions, often called kernel-based methods, have been studied extensively and applied to wide spectrum of pattern classification and function approximation problems. Basically, the main idea behind SVM is to find the largest margin hyperplane that separates two classes, among all the possible hyperplanes. The minimal distance from the separating hyperplane to the closest training example is called margin. Thus, the optimal hyperplane is the one providing the maximal margin, which represents the largest separation between the classes. This will be the line such that the distances from the closest point in each of the two groups will be farthest away. The training samples that lie on the margin are referred as support vectors, and conceptually are the most difficult data points to classify. Therefore, support vectors define the location of the separating hyperplane, being located at the boundary of their respective classes. See Figure B.1 to find a representation of support vectors and margin.

In the binary classification setting, let $((x_1, y_1)...(x_n, y_n))$ be the training dataset where x_i are the feature vectors representing the instances and $y_i \in \{-1, +1\}$ denote the labels of the instances. Support vector learning is the problem of finding a separating hyperplane that separates the positive examples from the negatives examples with the largest margin. The margin of the hyperplane is defined as the shortest distance between the positive and negative instances that are closest to the hyperplane. The intuition behind searching for the hyperplane with a large margin is that a hyperplane with the largest margin should be more resistant to noise than a hyperplane with a smaller margin. Supposing that all the training data satisfy the following constraints:

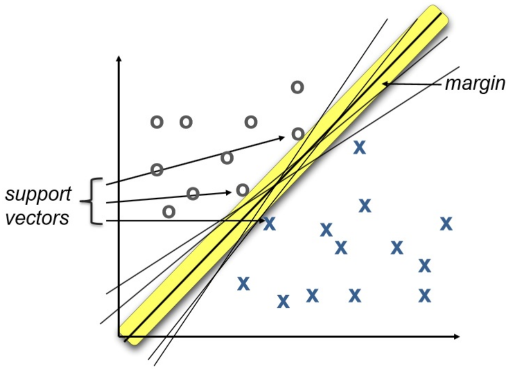


Figure B.1: The max-margin approach favored by Support Vector Machines.

$$w \cdot x_i + b \geq +1, \quad for \quad y_i = +1 \tag{B.1}$$

$$w \cdot x_i + b \leq -1, \quad for \quad y_i = -1 \tag{B.2}$$

where w is normal to the hyperplane, $|b|/||w||$ is the perpendicular distance from the hyperplane to the origin, and $||w||$ is the Euclidean norm of w. For the linearly separable case, the support vector algorithm looks for the separating hyperplane with largest margin, which can be formulated as follows:

$$y_i(w \cdot x_i + b - 1) \quad \geq 0 \quad \forall i \tag{B.3}$$

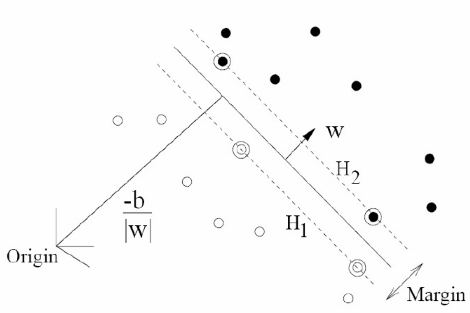


Figure B.2: Linear separating hyperplanes for the binary separable case. Circled exampled that lie on the hyperplane are called support vectors.

Let consider now that points in Eq. B.1 and Eq. B.2 lie on the hyperplanes H_1 and H_2 in Figure B.2, respectively:

$$H_1 : w \cdot x_i + b = 1 \tag{B.4}$$

$$H_2 : w \cdot x_i + b = -1 \tag{B.5}$$

with normal w and perpendicular distance from the origin $|1 - b|/\|w\|$ for the first case, and $|-1 - b|/\|w\|$ for the second case. Hence, the shortest distance from the separating hyperplane to the closest positive and negative examples is defined as $1/\|w\|$, and the margin is simply two times this distance, $2/\|w\|$. Thus, the maximum margin that separates the two classes can be constructed by solving the following primal optimization problem

$$minimize \quad \frac{1}{2}\|w\|^2 \tag{B.6}$$

subject to the constraints given by Eq. B.3. In other words, the margin is maximized, subject to the constraints that all training cases fall on either side of the support hyper-planes. The cases that lie on the hyperplane are called support vectors, since they support the hyper-planes and hence determine the solution to the problem. The primal problem can be solved by a quadratic program. However, it is not ready to be kernelised, because its dependence is not only on inner products between data-vectors.

A switch to Lagrangian formulation of the primal problem is done at this point mainly because of two reasons. First, the constraints are easier to handle. And second, the training data only appears as a dot product between vectors in this reformulation. This second characteristic of the Lagrangian reformulation is an essential property in order to generalize the procedure to the nonlinear case. Hence, positive Lagrange multipliers $\alpha_i, i = 1, ..., l$, for each of the inequality constraints in B.3 are introduced. The generalized Lagrangian function is then defined as:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i y_i (x_i w + b) + \sum_{i=1}^{l} \alpha_i \tag{B.7}$$

The goal is to minimize (B.7) with respect to w, b, and simultaneously require that the derivatives of $L(w, b, \alpha)$ with respect to all the α_i vanish, subjected to the constraints $\alpha_i \geq 0$.

B.0.2.1 Duality

Optimization problems can be converted to their dual form by differentiating the Lagrangian with regards to the original variables, solving the obtained results for those variables if possible, and substituting the resulting expression(s) back into the Lagrangian, thereby eliminating the variables.

Minimizing Eq. B.7 is a convex quadratic programming problem, because the objective function is itself convex, and points satisfying the constraints also form a convex set. In these cases, and only then, minimization and maximization can be interchanged, allowing to equivalently solve what is known as the “dual” problem. Duality of the problem is defined as maximizing $L(w, b, \alpha)$ subject to the constraints that the gradient of $L(w, b, \alpha)$ with respect to w and b vanish, and also subject to the constraints that the $\alpha_i \geq 0$. Forcing the gradient of $L(w, b, \alpha)$ with respect to w and b vanish give the conditions:

$$w = \sum_{i=1} \alpha_i y_i x_i \tag{B.8}$$

$$\sum_{i=1} \alpha_i y_i = 0 \tag{B.9}$$

Inserting this back into the Lagrangian formulation (Eq. B.7), the formulation of the dual problem becomes:

$$maximize \quad L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \tag{B.10}$$

which is subject to the constraints of B.9. The hyperplane whose weight vector $w^* = \sum_{i=1}^{n} y_i \alpha_i x_i$ solves this quadratic optimization problem is the maximal margin hyperplane with geometric margin $\lambda = \frac{1}{\|w\|}$.

The theory of duality guarantees that for convex problems, the dual problem becomes concave with an unique solution of the primal problem that corresponds to the unique solution of the dual problem. The important point of problem dualization is that the dual problem only depends on x_i through the inner product $x_i x_j$. A clear advantage is that the dual problem lends itself to kernelization, via the substitution $x_i x_j \longrightarrow k(x_i, x_j)$, while the primal problem does not.

B.0.2.2 The Karush-Kuhn-Tucker Conditions

The Karush-Kuhn-Tucker (KKT) conditions [195, 196] establish the requirements that need to be satisfied by an optimum solution to a general optimization problem. Given the primal problem in B.7, KKT conditions state that the solutions w^*, b^* and α^* should satisfy the following conditions (where i runs from 1 to the number of training points and v from 1 to the dimension of the data d)

$$\frac{\partial L(w^*, b^*, \alpha^*)}{\partial w_v} = w_v - \sum_i \alpha_i y_i x_{iv} = 0 \quad v = 1, ..., d \tag{B.11}$$

$$\frac{\partial L(w^*, b^*, \alpha^*)}{\partial b} = -\sum_i \alpha_i y_i = 0 \tag{B.12}$$

$$y_i(x_i \cdot w + b) - 1 \geq 0, \qquad \forall i \tag{B.13}$$

$$\alpha_i \geq 0, \qquad \forall i \tag{B.14}$$

$$\alpha_i(y_i(w \cdot x_i + b) - 1) = 0, \qquad \forall i \tag{B.15}$$

Since the problem for SVM is convex, KKT conditions are necessary and sufficient for w^*, b^* and α^* to be a solution [197]. Hence, solving the SVM problem is equivalent to finding a solution to the KKT conditions. The first KKT condition (Eq. B.11) defines the optimal hyperplane as a linear combination of the vectors in the training set:

$$w^* = \sum_i \alpha_i^* y_i x_i \tag{B.16}$$

In the other hand, the second KKT condition (Eq. B.12) requires that the α_i coefficients of the training instances should satisfy:

$$\sum_{i=1}^{n} \alpha_i^* y_i = 0, \qquad \alpha_i^* \geq 0 \tag{B.17}$$

As an application of the KKT conditions, the decision function that can be used to classify future test cases is defined as:

$$f(x) = w^T x_i + b = \sum_i \alpha_i y_i x_i^T x + b \tag{B.18}$$

where the sign of the decision function determines the predicted classification of x.

The most important conclusions are that, first, this function $f(\cdot)$ can be expressed solely in terms of inner products $x_i^T x_i$, that can be later replaced with kernel matrices $k(x_i, x_j)$ to move to a higher dimensional non-linear spaces. Second, only support vectors are needed to express the solution w.

B.0.2.3 The Non-Separable Case

However, most of the real data is not linearly separable and, even in cases where the data is linearly separable, SVM may overfit to the training data in its search for the hyperplane that completely separates all of the instances of both classes. Sometimes, even if a curved decision boundary is possible, exactly separating the data is probably not desirable: if the data has noise and outliers, a smooth decision boundary that ignores a few data points is better than one that loops around the outliers. Therefore, linear separation may present a high sensivity to ourliers. To address these problems, the concept of "soft margin" were introduced in SVM by [141]. The basic is to relax the constraints in B.1 and B.2, only when necessary, via the introduction a further cost in the primal objective function (Eq. B.6). This can be done by introducing positive "slack variables" ξ_i in the constraints. With the addition of the "slack variables", the modified relaxed constraints become

$$w \cdot x_i + b \geq +1 - \xi_i, \quad for \quad y_i = +1 \tag{B.19}$$

$$w \cdot x_i + b \leq -1 - \xi_i, \quad for \quad y_i = -1 \tag{B.20}$$

$$\xi_i \geq 0 \tag{B.21}$$

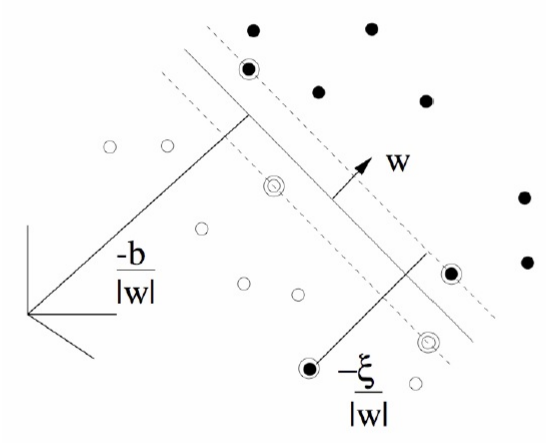


Figure B.3: Soft Margin SVM.

The introduction of "slack variables" allows for violations of the constraint, i.e. permits some instances to lie inside the margin or even cross further among the instances of the opposite class (Fig. B.3). To avoid arbitrarily large values for ξ_i that would cause the SVM to obtain trivial and suboptimal solutions, the relaxation must be constrained. Adding the "slack variables" in the objective function (Eq. B.6) allows to control the relaxation. The new primal "relaxed" problem this becomes

$$minimize \quad L_P = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \tag{B.22}$$

subject to B.19, B.20 and B.21. The penalty parameter $C > 0$ controls the trade-off between the penalty and margin, i.e. specifies the misclassification penalty. Its value is tuned by the user and it is based on the classification problem and dataset characteristics. Small C values allow constraints to be easily ignored, i.e. large margin, while large C values makes constraints hard to ignore, i.e. narrow margin.

As in the linear case, the problem is switched to a Lagrangian formulation, leading to

$$L(w, b, \xi, \alpha, \mu) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_{i=1}^{N} \alpha_i[y_i(w^T x_i - b) - 1 + \xi_i] - \sum_{i=1}^{N} \mu_i \xi_i \tag{B.23}$$

Derived KKT conditions are defined as

$$\frac{\partial L(w^*, b^*, \xi^*, \alpha^*, \mu^*)}{\partial w_v} = w_v - \sum_i \alpha_i y_i x_{iv} = 0 \tag{B.24}$$

$$\frac{\partial L(w^*, b^*, \xi^*, \alpha^*, \mu^*)}{\partial b} = -\sum_i \alpha_i y_i = 0 \tag{B.25}$$

$$\frac{\partial L(w^*, b^*, \xi^*, \alpha^*, \mu^*)}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \tag{B.26}$$

$$y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0 \tag{B.27}$$

$$\xi_i \geq 0 \tag{B.28}$$

$$\alpha_i \geq 0 \tag{B.29}$$

$$\mu_i \geq 0 \tag{B.30}$$

$$\alpha_i y_i(x_i \cdot w + b) - 1 + \xi_i = 0 \tag{B.31}$$

$$\mu_i \xi_i \geq 0 \tag{B.32}$$

Conveniently converting to the dual problem, and using the KKT equations, the SVM problem can be then efficiently solved, as well as it becomes readily kernelized. The dual formulation is then defined as

$$maximize \quad L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \tag{B.33}$$

$$subjectto \qquad 0 \leq \alpha_i \leq C, \qquad and \qquad \sum_i \alpha_i y_i = 0. \tag{B.34}$$

The solution is again given by

$$w = \sum_{i=1}^{N_S} \alpha_i y_i x_i \tag{B.35}$$

where N_S is the number of supported vectors. This solutions is practically the same than in the linear case, but with an extra constraint on the multipliers α_i which have now an upper bound of C.

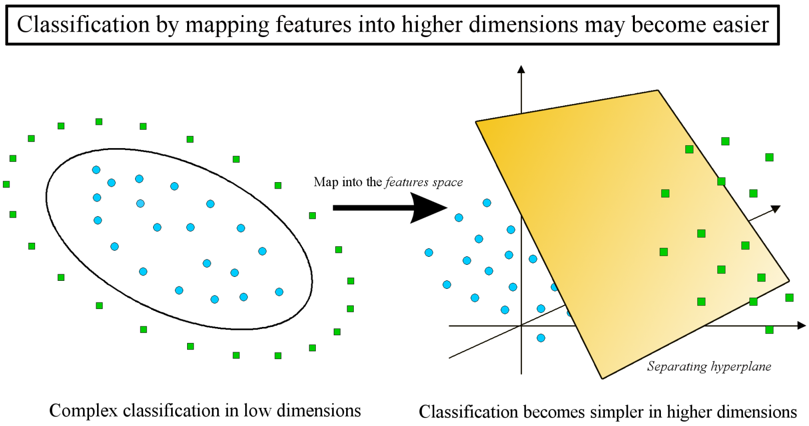
B.0.2.4 Non-linear Support Vector Machines

The power of SVMs can be fully realized when linear SVMs are extended to allow more general decision surfaces. One of the benefits of Lagrangian reformulation of the SVM problem is that the training data appears as a dot product between vectors (Section X). This advantage can be exploited by using the kernel trick, which allows SVM to form non-linear boundaries.

In the dual problem in B.33, the dot product can be replaced with the new kernel function K,

$$maximize \quad L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{B.36}$$

subject to the conditions defined in B.34.


Classification by mapping features into higher dimensions may become easier

Complex classification in low dimensions

Classification becomes simpler in higher dimensions

Figure B.4: Effect of the kernel transformation. Data is not linearly separable in (a). Mapping features into a higher dimensionality (b) may make the classification possible.

B.0.2.5 Kernel selection and parameters tuning

As explained in previous section, to map input features into higher dimensionality spaces several kernels can be used, according to the nature of the data. This is done via some mapping $\Phi(x)$ and then, construction of a separating hyperplane with maximum margin is done in the input space (Figure B.4). As shown in B.4, the linear decision function in the features space corresponds to a non-linear decision boundary in the original input space. Typical kernels' choices are ([148]):

- Linear Kernel: $K(x, y) = \langle x, y \rangle$
- Polynomial Kernel: $K(x, y) = (\langle x, y \rangle)^2$
- RBF Kernel: $K(x, y) = exp(-\gamma \|x - y\|^2)$
- Sigmoid Kernel $K(x, y) = tanh(\gamma \langle x, y \rangle - theta)$
- Histogram Intersection Kernel $K(x, y) = |x - y|$

Each kernel function listed above has its own properties and unique response to handle a variety of data. Employing a sigmoid kernel function in a SVM model is equivalent to use a two-layer perceptron neural network [143]. If RBF kernel is used instead, the model approximately behaves like a radial basis function neural network, where the feature space is in an infinite dimension. Therefore, selection of a proper kernel function is required to perform optimal classification tasks with SVM models. Selection of the convenient kernel function is, or should be, based on the requirements of the classification task.

B.1 SVM Parameter Setting

First choice to make when working with SVM is the kernel to be used. Despite the several kernels proposed to map features into a higher dimension, Radial Basis Function (RBF) kernels are one of the most used kernels to separate data in SVM classifiers in complex classification environments. Some previous works have found that RBF kernel generally provides better classification accuracy than many other kernel functions [198]. This kernel non-linearly maps samples into a higher dimensional space. That means that RBF kernel can handle the cases when the relation between class labels and attributes is nonlinear. Second reason to use this kernel is the number of hyperparameters which influences the complexity of model selection, which is lower than in other non-linear kernels, such as the polynomial kernel. Consider two samples $x_i = [x_{i1}, x_{i2}, ..., x_{id}]^T$ and $x_j = [x_{j1}, x_{j2}, ..., x_{jd}]^T$. The RBF kernel is then defined by:

$$K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right), \quad \gamma > 0 \tag{B.37}$$

where γ is the width of the Gaussian.

There are two parameters that can be tuned in the RBF kernel and which depend on the input data: C and γ. While C controls the cost of misclassification on the training data, γ is the parameter of the kernel to handle non-linear classification. A large C value will provide a low bias and high variance, because misclassification cost is highly penalized, i.e. hard margin. Contrary, a small C value makes the cost of misclassification low, i.e. soft margin, giving a higher bias and lower variance. To “raise” the points used in the RBF kernel, γ controls the shape of the “peaks" where the points are raised (Fig. B.5). A large γ will give a pointed bump in the higher dimensions, while a small γ will give a softer, broader bump. This is translated into low bias and high variance with a large γ value and higher bias and low variance for lower γ values. Thus, a γ overestimation will produce an almost linear behavior in the exponential and the higher-dimensional projection would start to lose its non-linear power. However, an underestimated γ value will produce a lack on regularization, making the decision boundary highly sensitive to noise in the training data. Therefore, some kind of model selection, i.e. parameter search, must be done for these two parameters. The goal of this search is to identify good C and γ values so that the classifier can accurately predict unknown data, i.e. testing data. Since performing a fully grid search may become time consuming, a coarse grid search is often initially conducted. After identifying the best region on the grid, a finer grid search on that region can be performed.

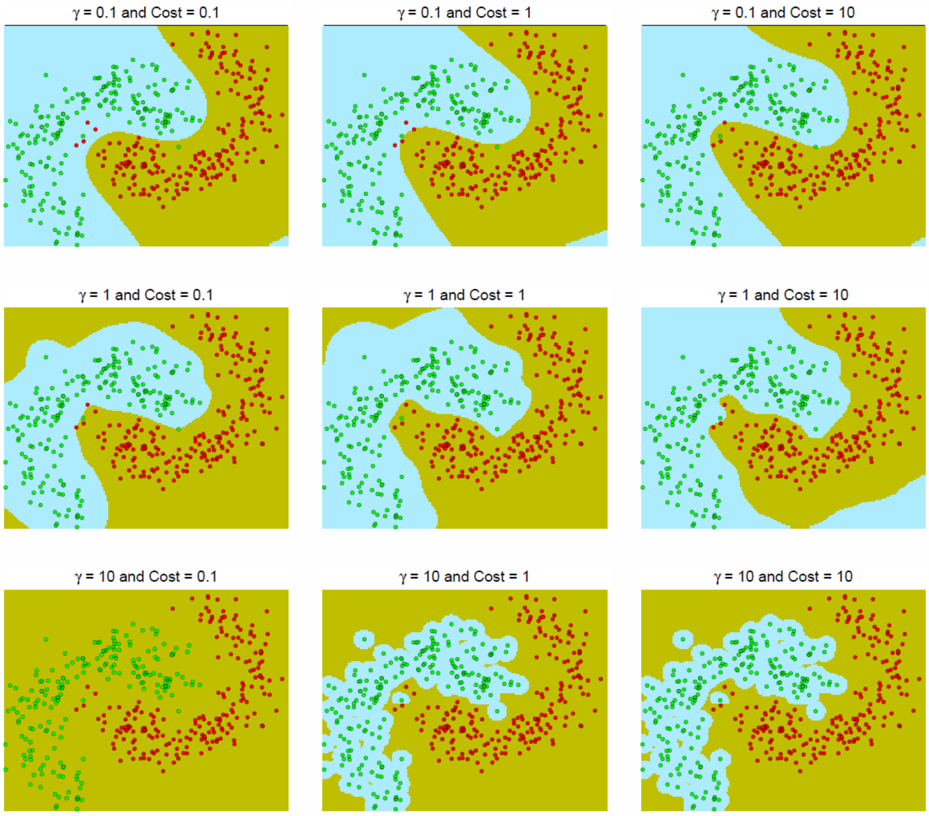


Figure B.5: Decision boundaries for a banana shaped dataset generated by SVM with a RBF kernel for different C and γ values.

List of Figures

List of Tables

Bibliography

[1] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al.. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer;; 2013. Available from: http://globocan.iarc.fr(accessedDecember16,2014). (Cited on pages 1 and 5.)

[2] Albreht T, McKee M, Alexe DM, Coleman MP, Martin-Moreno JM. Making progress against cancer in Europe in 2008. European Journal of Cancer. 2008;44(10):1451–1456. (Cited on pages 1, 5 and 178.)

[3] Thun MJ, DeLancey JO, Center MM, Jemal A, Ward EM. The global burden of cancer: priorities for prevention. Carcinogenesis. 2010;31(1):100–110. (Cited on pages 1 and 5.)

[4] Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA: a cancer journal for clinicians. 2014;64(1):9–29. (Cited on pages 1, 5 and 178.)

[5] Bondiau PY, Malandain G, Chanalet S, Marcy PY, Habrand JL, Fauchon F, et al. Atlas-based automatic segmentation of MR images: validation study on the brainstem in radiotherapy context. International Journal of Radiation Oncology* Biology* Physics. 2005;61(1):289–298. (Cited on pages 1, 22, 26, 27, 54, 55, 164, 165 and 166.)

[6] Deeley M, Chen A, Datteri R, Noble J, Cmelak A, Donnelly E, et al. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. Physics in medicine and biology. 2011;56(14):4557. (Cited on pages 1, 26, 110, 165, 166 and 168.)

[7] Cohen K. Brain Tumors; Leaving the Garden of Eden. Wiley Online Library; 2005. (Cited on page 5.)

[8] Mesulam M, et al. Principles of behavioral and cognitive neurology . Oxford University Press; 2000. (Cited on page 5.)

[9] Ward J. DNA damage produced by ionizing radiation in mammalian cells: identities, mechanisms of formation, and reparability. Progress in nucleic acid research and molecular biology. 1988;35:95–125. (Cited on pages 6 and 178.)

[10] Joiner MC, van der Kogel A. Basic Clinical Radiobiology Fourth Edition. CRC Press; 2009. (Cited on page 7.)

[11] Floyd SR, Kasper EM, Uhlmann EJ, Fonkem E, Wong ET, Mahadevan A. Hypofractionated radiotherapy and stereotactic boost with concurrent and

adjuvant temozolamide for glioblastoma in good performance status elderly patients–early results of a phase II trial. Frontiers in oncology. 2012;2. (Cited on page 8.)

[12] Combs SE, Widmer V, Thilmann C, Hof H, Debus J, Schulz-Ertner D. Stereotactic radiosurgery (SRS). Cancer. 2005;104(10):2168–2173. (Cited on page 9.)

[13] Barnett GH, Linskey ME, Adler JR, Cozzens JW, Friedman WA, Heilbrun MP, et al. Stereotactic radiosurgery–an organized neurosurgery-sanctioned definition. Journal of neurosurgery. 2007;106(1):1. (Cited on page 9.)

[14] De Salles AA, Gorgulho A, Selch M, De Marco J, Agazaryan N. Radiosurgery from the brain to the spine: 20 years experience. Springer; 2008. (Cited on page 9.)

[15] Grimm J, LaCouture T, Croce R, Yeo I, Zhu Y, Xue J. Dose tolerance limits and dose volume histogram evaluation for stereotactic body radiotherapy. Journal of Applied Clinical Medical Physics. 2011;12(2). (Cited on page 18.)

[16] Hunt MA, Zelefsky MJ, Wolden S, Chui CS, LoSasso T, Rosenzweig K, et al. Treatment planning and delivery of intensity-modulated radiation therapy for primary nasopharynx cancer. International Journal of Radiation Oncology* Biology* Physics. 2001;49(3):623–632. (Cited on page 18.)

[17] Narayana A, Yamada J, Berry S, Shah P, Hunt M, Gutin PH, et al. Intensity-modulated radiotherapy in high-grade gliomas: clinical and dosimetric results. International Journal of Radiation Oncology* Biology* Physics. 2006;64(3):892–897. (Cited on page 18.)

[18] Bhandare N, Jackson A, Eisbruch A, Pan CC, Flickinger JC, Antonelli P, et al. Radiation therapy and hearing loss. International journal of radiation oncology, biology, physics. 2010;76(3 Suppl):S50. (Cited on page 18.)

[19] Timmerman RD. An Overview of Hypofractionation and Introduction to This Issue of Seminars in Radiation Oncology. In: Seminars in radiation oncology. vol. 18. WB Saunders; 2008. p. 215–222. (Cited on page 18.)

[20] Sharma MS, Kondziolka D, Khan A, Kano H, Niranjan A, Flickinger JC, et al. Radiation tolerance limits of the brainstem. Neurosurgery. 2008;63(4):728–733. (Cited on page 18.)

[21] Mould RF. Robotic radiosurgery. CyberKnife Society Press; 2005. (Cited on page 18.)

[22] Massager N, Nissim O, Delbrouck C, Delpierre I, Devriendt D, Desmedt F, et al. Irradiation of cochlear structures during vestibular schwannoma radiosurgery and associated hearing outcome. 2007;. (Cited on page 18.)

[23] Romanelli P, Muacevic A, Striano S. Radiosurgery for hypothalamic hamartomas. 2008;. (Cited on page 18.)

[24] Lee M, Kalani MYS, Cheshier S, Gibbs IC, Adler Jr JR, Chang SD. Radiation therapy and CyberKnife radiosurgery in the management of craniopharyngiomas. 2008;. (Cited on page 18.)

[25] Stafford SL, Pollock BE, Leavitt JA, Foote RL, Brown PD, Link MJ, et al. A study on the radiation tolerance of the optic nerves and chiasm after stereotactic radiosurgery. International Journal of Radiation Oncology* Biology* Physics. 2003;55(5):1177–1181. (Cited on page 18.)

[26] Sheehan JP, Gerszten P. Controversies in Stereotactic Radiosurgery: Best Evidence Recommendations. Thieme; 2013. (Cited on pages 19 and 179.)

[27] Rees J. Advances in magnetic resonance imaging of brain tumours. Current opinion in neurology. 2003;16(6):643–650. (Cited on page 19.)

[28] Whitfield GA, Price P, Price GJ, Moore CJ. Automated delineation of radiotherapy volumes: are we going in the right direction? The British journal of radiology. 2013;86(1021):20110718–20110718. (Cited on pages 21 and 179.)

[29] Yamamoto M, Nagata Y, Okajima K, Ishigaki T, Murata R, Mizowaki T, et al. Differences in target outline delineation from CT scans of brain tumours using different methods and different observers. Radiotherapy and oncology. 1999;50(2):151–156. (Cited on page 21.)

[30] Mazzara GP, Velthuizen RP, Pearlman JL, Greenberg HM, Wagner H. Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation. International Journal of Radiation Oncology* Biology* Physics. 2004;59(1):300–312. (Cited on page 21.)

[31] Nelms BE, Tomé WA, Robinson G, Wheeler J. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. International Journal of Radiation Oncology* Biology* Physics. 2012;82(1):368–378. (Cited on page 21.)

[32] D'Haese PFD, Duay V, Li R, du Bois d'Aische A, Merchant TE, Cmelak AJ, et al. Automatic segmentation of brain structures for radiation therapy planning. In: Medical Imaging 2003. International Society for Optics and Photonics; 2003. p. 517–526. (Cited on pages 22, 26 and 180.)

[33] Pham DL, Xu C, Prince JL. Current methods in medical image segmentation 1. Annual review of biomedical engineering. 2000;2(1):315–337. (Cited on page 24.)

[34] Xuan J, Adali T, Wang Y. Segmentation of magnetic resonance brain image: integrating region growing and edge detection. In: Image Processing, 1995.

Proceedings., International Conference on. vol. 3. IEEE; 1995. p. 544–547. (Cited on pages 25 and 181.)

[35] Balafar MA, Ramli AR, Saripan MI, Mashohor S. Review of brain MRI image segmentation methods. Artificial Intelligence Review. 2010;33(3):261–274. (Cited on page 26.)

[36] Senthilkumaran N, Rajesh R. Brain image segmentation. International Journal of Wisdom Based Computing. 2011;1(3):14–18. (Cited on page 26.)

[37] Lee CH, Schmidt M, Murtha A, Bistritz A, Sander J, Greiner R. Segmenting brain tumors with conditional random fields and support vector machines. In: Computer vision for biomedical image applications. Springer; 2005. p. 469–478. (Cited on pages 26 and 181.)

[38] Norman KA. How hippocampus and cortex contribute to recognition memory: revisiting the complementary learning systems model. Hippocampus. 2010;20(11):1217–1227. (Cited on page 26.)

[39] Laakso M, Partanen K, Riekkinen P, Lehtovirta M, Helkala EL, Hallikainen M, et al. Hippocampal volumes in Alzheimer's disease, Parkinson's disease with and without dementia, and in vascular dementia An MRI study. Neurology. 1996;46(3):678–681. (Cited on page 26.)

[40] Ghanei A, Soltanian-Zadeh H, Windham JP. Segmentation of the hippocampus from brain MRI using deformable contours. Computerized Medical Imaging and Graphics. 1998;22(3):203–216. (Cited on pages 26, 38, 40, 54 and 55.)

[41] Shen D, Moffat S, Resnick SM, Davatzikos C. Measuring size and shape of the hippocampus in MR images using a deformable shape model. Neuroimage. 2002;15(2):422–434. (Cited on pages 26, 54 and 56.)

[42] Hult R. Grey-level morphology combined with an artificial neural networks approach for multimodal segmentation of the Hippocampus. In: Image Analysis and Processing, 2003. Proceedings. 12th International Conference on. IEEE; 2003. p. 277–282. (Cited on pages 26, 41, 47, 54 and 185.)

[43] Morra JH, Tu Z, Apostolova LG, Green AE, Toga AW, Thompson PM. Automatic subcortical segmentation using a contextual model. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008. Springer; 2008. p. 194–201. (Cited on pages 26, 41, 49, 54, 56, 185 and 186.)

[44] Artaechevarria X, Munoz-Barrutia A, Ortiz-de Solórzano C. Combination strategies in multi-atlas image segmentation: Application to brain MR data. Medical Imaging, IEEE Transactions on. 2009;28(8):1266–1277. (Cited on pages 26, 27, 30, 31 and 54.)

[45] Collins DL, Pruessner JC. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. Neuroimage. 2010;52(4):1355–1366. (Cited on pages 26, 27, 30, 54 and 55.)

[46] Coupé P, Manjón JV, Fonov V, Pruessner J, Robles M, Collins DL. Nonlocal patch-based label fusion for hippocampus segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010. Springer; 2010. p. 129–136. (Cited on pages 26, 27, 30, 31, 54 and 55.)

[47] Morra JH, Tu Z, Apostolova LG, Green AE, Toga AW, Thompson PM. Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation. Medical Imaging, IEEE Transactions on. 2010;29(1):30–43. (Cited on pages 26, 41, 49, 50, 54, 56, 185 and 186.)

[48] Hu S, Coupé P, Pruessner JC, Collins DL. Appearance-based modeling for segmentation of hippocampus and amygdala using multi-contrast MR imaging. NeuroImage. 2011;58(2):549–559. (Cited on pages 26, 33, 54, 55 and 183.)

[49] Khan AR, Cherbuin N, Wen W, Anstey KJ, Sachdev P, Beg MF. Optimal weights for local multi-atlas fusion using supervised learning and dynamic information (SuperDyn): Validation on hippocampus segmentation. NeuroImage. 2011;56(1):126–139. (Cited on pages 26, 27, 29, 30, 31, 54 and 55.)

[50] Kim M, Wu G, Li W, Wang L, Son YD, Cho ZH, et al. Segmenting hippocampus from 7.0 Tesla MR images by combining multiple atlases and auto-context models. In: Machine Learning in Medical Imaging. Springer; 2011. p. 100–108. (Cited on pages 26, 27, 30, 31, 54 and 55.)

[51] Zhao S, Zhang D, Song X, Tan W. Segmentation of hippocampus in MRI images based on the improved level set. In: Computational Intelligence and Design (ISCID), 2011 Fourth International Symposium on. vol. 1. IEEE; 2011. p. 123–126. (Cited on pages 26, 38 and 54.)

[52] Cardoso MJ, Leung K, Modat M, Keihaninejad S, Cash D, Barnes J, et al. STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation. Medical image analysis. 2013;17(6):671–684. (Cited on pages 26, 30, 31, 54 and 55.)

[53] Kwak K, Yoon U, Lee DK, Kim GH, Seo SW, Na DL, et al. Fully-automated approach to hippocampus segmentation using a graph-cuts algorithm combined with atlas-based segmentation and morphological opening. Magnetic resonance imaging. 2013;31(7):1190–1196. (Cited on pages 26, 54 and 55.)

[54] Wang H, Suh JW, Das SR, Pluta JB, Craige C, Yushkevich PA. Multi-atlas segmentation with joint label fusion. Pattern Analysis and Machine Intelli-

gence, IEEE Transactions on. 2013;35(3):611–623. (Cited on pages 26, 27, 29, 30, 31, 54 and 56.)

[55] Zarpalas D, Gkontra P, Daras P, Maglaveras N. Hippocampus segmentation through gradient based reliability maps for local blending of ACM energy terms. In: Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on. IEEE; 2013. p. 53–56. (Cited on pages 26 and 54.)

[56] McIntosh C, Hamarneh G. Medial-based deformable models in nonconvex shape-spaces for medical image segmentation. Medical Imaging, IEEE Transactions on. 2012;31(1):33–50. (Cited on pages 26, 38, 39 and 54.)

[57] Leventon ME, Grimson WEL, Faugeras O. Statistical shape influence in geodesic active contours. In: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on. vol. 1. IEEE; 2000. p. 316–323. (Cited on pages 26, 31, 38, 40, 54 and 55.)

[58] Olveres J, Nava R, Escalante-Ramírez B, Cristóbal G, García-Moreno CM. Midbrain volume segmentation using active shape models and LBPs. In: SPIE Optical Engineering+ Applications. International Society for Optics and Photonics; 2013. p. 88561F–88561F. (Cited on pages 26, 33, 35, 36, 54, 56 and 183.)

[59] Cootes TF, Beeston C, Edwards GJ, Taylor CJ. A unified framework for atlas matching using active appearance models. In: Information Processing in Medical Imaging. Springer; 1999. p. 322–333. (Cited on pages 26, 33, 36, 40, 51, 54, 55 and 183.)

[60] Duchesne S, Pruessner J, Collins D. Appearance-based segmentation of medial temporal lobe structures. Neuroimage. 2002;17(2):515–531. (Cited on pages 26, 33, 36, 54, 55 and 183.)

[61] Bailleul J, Ruan S, Bloyet D, Romaniuk B. Segmentation of anatomical structures from 3D brain MRI using automatically-built statistical shape models. In: Image Processing, 2004. ICIP'04. 2004 International Conference on. vol. 4. IEEE; 2004. p. 2741–2744. (Cited on pages 26, 33, 35, 54 and 183.)

[62] Babalola KO, Cootes TF, Twining CJ, Petrovic V, Taylor C. 3D brain segmentation using active appearance models and local regressors. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008. Springer; 2008. p. 401–408. (Cited on pages 26, 33, 37, 54, 55 and 183.)

[63] Tu Z, Narr KL, Dollár P, Dinov I, Thompson PM, Toga AW. Brain anatomical structure segmentation by hybrid discriminative/generative models. Medical Imaging, IEEE Transactions on. 2008;27(4):495–508. (Cited on pages 26, 33, 35, 54, 56 and 183.)

[64] Hu S, Coupé P, Pruessner JC, Collins DL. Nonlocal regularization for active appearance model: Application to medial temporal lobe segmentation. Human brain mapping. 2014;35(2):377–395. (Cited on pages 26, 33, 54, 55 and 183.)

[65] Heckemann RA, Hajnal JV, Aljabar P, Rueckert D, Hammers A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage. 2006;33(1):115–126. (Cited on pages 26, 27, 30, 54 and 55.)

[66] Wu M, Rosano C, Lopez-Garcia P, Carter CS, Aizenstein HJ. Optimum template selection for atlas-based segmentation. NeuroImage. 2007;34(4):1612–1618. (Cited on pages 26, 27, 29, 30, 54 and 56.)

[67] Aljabar P, Heckemann RA, Hammers A, Hajnal JV, Rueckert D. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. Neuroimage. 2009;46(3):726–738. (Cited on pages 26, 27, 29, 30, 33, 54 and 55.)

[68] Lötjönen JM, Wolz R, Koikkalainen JR, Thurfjell L, Waldemar G, Soininen H, et al. Fast and robust multi-atlas segmentation of brain magnetic resonance images. Neuroimage. 2010;49(3):2352–2365. (Cited on pages 26, 27, 29, 54 and 56.)

[69] Asman AJ, Landman BA. Non-local statistical label fusion for multi-atlas segmentation. Medical image analysis. 2013;17(2):194–208. (Cited on pages 26, 27, 30, 31, 54 and 55.)

[70] Székely G, Kelemen A, Brechbühler C, Gerig G. Segmentation of 3D objects from MRI volume data using constrained elastic deformations of flexible Fourier surface models. In: Computer Vision, Virtual Reality and Robotics in Medicine. Springer; 1995. p. 495–505. (Cited on pages 26, 38, 39 and 54.)

[71] Tsai A, Wells W, Tempany C, Grimson E, Willsky A. Mutual information in coupled multi-shape model for medical image segmentation. Medical Image Analysis. 2004;8(4):429–445. (Cited on pages 26, 38, 40 and 54.)

[72] Yang J, Duncan JS. 3D image segmentation of deformable objects with joint shape-intensity prior models using level sets. Medical Image Analysis. 2004;8(3):285–294. (Cited on pages 26, 38, 40, 54 and 56.)

[73] Magnotta VA, Heckel D, Andreasen NC, Cizadlo T, Corson PW, Ehrhardt JC, et al. Measurement of Brain Structures with Artificial Neural Networks: Two-and Three-dimensional Applications 1. Radiology. 1999;211(3):781–790. (Cited on pages 26, 41, 47, 48, 54, 56, 85 and 185.)

[74] Pierson R, Corson PW, Sears LL, Alicata D, Magnotta V, O'Leary D, et al. Manual and semiautomated measurement of cerebellar subregions on MR im-

ages. Neuroimage. 2002;17(1):61–76. (Cited on pages 26, 41, 47, 48, 54, 56 and 185.)

[75] Golland P, Grimson WEL, Shenton ME, Kikinis R. Detection and analysis of statistical differences in anatomical shape. Medical image analysis. 2005;9(1):69–86. (Cited on pages 26, 41, 49, 50, 54, 55, 185 and 186.)

[76] Powell S, Magnotta VA, Johnson H, Jammalamadaka VK, Pierson R, Andreasen NC. Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. Neuroimage. 2008;39(1):238–247. (Cited on pages 26, 41, 47, 48, 49, 50, 52, 54, 56, 85, 185 and 186.)

[77] Gensheimer M, Cmelak A, Niermann K, Dawant BM. Automatic delineation of the optic nerves and chiasm on CT images. In: Medical Imaging. International Society for Optics and Photonics; 2007. p. 651216–651216. (Cited on page 26.)

[78] Isambert A, Dhermain F, Bidault F, Commowick O, Bondiau PY, Malandain G, et al. Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context. Radiotherapy and oncology. 2008;87(1):93–99. (Cited on pages 26, 164, 165 and 166.)

[79] Noble JH, Dawant BM. An atlas-navigated optimal medial axis and deformable model algorithm (NOMAD) for the segmentation of the optic nerves and chiasm in MR and CT images. Medical image analysis. 2011;15(6):877–884. (Cited on pages 26, 163, 165 and 166.)

[80] Conson M, Cella L, Pacelli R, Comerci M, Liuzzi R, Salvatore M, et al. Automated delineation of brain structures in patients undergoing radiotherapy for primary brain tumors: From atlas to dose–volume histograms. Radiotherapy and Oncology. 2014;112(3):326–331. (Cited on page 26.)

[81] Panda S, Asman AJ, DeLisi MP, Mawn LA, Galloway RL, Landman BA. Robust optic nerve segmentation on clinically acquired CT. In: SPIE Medical Imaging. International Society for Optics and Photonics; 2014. p. 90341G–90341G. (Cited on pages 27, 30, 31, 54, 56, 165 and 166.)

[82] Cabezas M, Oliver A, Lladó X, Freixenet J, Bach Cuadra M. A review of atlas-based segmentation for magnetic resonance brain images. Computer methods and programs in biomedicine. 2011;104(3):e158–e177. (Cited on page 29.)

[83] Hill DL, Batchelor PG, Holden M, Hawkes DJ. Medical image registration. Physics in medicine and biology. 2001;46(3):R1. (Cited on page 29.)

[84] Zitova B, Flusser J. Image registration methods: a survey. Image and vision computing. 2003;21(11):977–1000. (Cited on page 29.)

[85] Toga AW, Thompson PM. The role of image registration in brain mapping. Image and vision computing. 2001;19(1):3–24. (Cited on pages 29 and 182.)

[86] Rohlfing T, Brandt R, Menzel R, Maurer Jr CR. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage. 2004;21(4):1428–1442. (Cited on pages 29, 30 and 33.)

[87] Han X, Hoogeman MS, Levendag PC, Hibbard LS, Teguh DN, Voet P, et al. Atlas-based auto-segmentation of head and neck CT images. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008. Springer; 2008. p. 434–441. (Cited on pages 29 and 30.)

[88] Commowick O, Akhondi-Asl A, Warfield SK. Estimating a reference standard segmentation with spatially varying performance parameters: Local MAP STAPLE. Medical Imaging, IEEE Transactions on. 2012;31(8):1593–1606. (Cited on pages 30 and 31.)

[89] Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. Medical Imaging, IEEE Transactions on. 2004;23(7):903–921. (Cited on pages 30 and 31.)

[90] Yezzi A, Zöllei L, Kapur T. A variational framework for joint segmentation and registration. In: Mathematical Methods in Biomedical Image Analysis, 2001. MMBIA 2001. IEEE Workshop on. IEEE; 2001. p. 44–51. (Cited on page 31.)

[91] Paragios N, Rousson M, Ramesh V. Knowledge-based registration & segmentation of the left ventricle: a level set approach. In: Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on. IEEE; 2002. p. 37–42. (Cited on pages 31 and 32.)

[92] Wyatt PP, Noble JA. MAP MRF joint segmentation and registration of medical images. Medical Image Analysis. 2003;7(4):539–552. (Cited on page 31.)

[93] Wang F, Vemuri BC. Simultaneous registration and segmentation of anatomical structures from brain MRI. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005. Springer; 2005. p. 17–25. (Cited on pages 31 and 32.)

[94] Pohl KM, Fisher J, Grimson WEL, Kikinis R, Wells WM. A Bayesian model for joint segmentation and registration. NeuroImage. 2006;31(1):228–239. (Cited on page 31.)

[95] Wu G, Wang L, Gilmore J, Lin W, Shen D. Joint segmentation and registration for infant brain images. In: Medical Computer Vision: Algorithms for Big Data. Springer; 2014. p. 13–21. (Cited on pages 31 and 32.)

[96] Gooya A, Pohl KM, Bilello M, Biros G, Davatzikos C. Joint segmentation and deformable registration of brain scans guided by a tumor growth model. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011. Springer; 2011. p. 532–540. (Cited on page 31.)

[97] Parisot S, Duffau H, Chemouny S, Paragios N. Joint tumor segmentation and dense deformable registration of brain MR images. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012. Springer; 2012. p. 651–658. (Cited on page 31.)

[98] Daisne JF, Blumhofer A. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. Radiation Oncology. 2013;8(1):154. (Cited on pages 33 and 182.)

[99] Cootes TF, Taylor CJ, Cooper DH, Graham J. Training models of shape from sets of examples. In: BMVC92. Springer; 1992. p. 9–18. (Cited on pages 33, 34, 35 and 183.)

[100] Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models-their training and application. Computer vision and image understanding. 1995;61(1):38–59. (Cited on pages 33, 34, 35, 51 and 183.)

[101] Brejl M, Sonka M. Object localization and border detection criteria design in edge-based image segmentation: automated learning from examples. Medical Imaging, IEEE Transactions on. 2000;19(10):973–985. (Cited on pages 33, 35, 54 and 183.)

[102] Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. IEEE Transactions on pattern analysis and machine intelligence. 2001;23(6):681–685. (Cited on pages 33, 35, 36, 40 and 183.)

[103] van Ginneken B, de Bruijne M, Loog M, Viergever MA. Interactive shape models. In: Medical Imaging 2003. International Society for Optics and Photonics; 2003. p. 1206–1216. (Cited on pages 33, 35, 37 and 183.)

[104] Pitiot A, Delingette H, Thompson PM, Ayache N. Expert knowledge-guided segmentation system for brain MRI. NeuroImage. 2004;23:S85–S96. (Cited on pages 33, 35, 54, 56 and 183.)

[105] Zhao Z, Aylward SR, Teoh EK. A novel 3D partitioned active shape model for segmentation of brain MR images. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005. Springer; 2005. p. 221–228. (Cited on pages 33, 35, 36, 54 and 183.)

[106] Koikkalainen J, Tolli T, Lauerma K, Antila K, Mattila E, Lilja M, et al. Methods of artificial enlargement of the training set for statistical shape models. Medical Imaging, IEEE Transactions on. 2008;27(11):1643–1654. (Cited on pages 33, 36 and 183.)

[107] Rao A, Aljabar P, Rueckert D. Hierarchical statistical shape analysis and prediction of sub-cortical brain structures. Medical image analysis. 2008;12(1):55–68. (Cited on pages 33, 35, 54, 56 and 183.)

[108] Heimann T, Meinzer HP. Statistical shape models for 3D medical image segmentation: A review. Medical image analysis. 2009;13(4):543–563. (Cited on pages 33, 34, 35 and 183.)

[109] Babalola K, Cootes T. Using parts and geometry models to initialise Active Appearance Models for automated segmentation of 3D medical images. In: Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on. IEEE; 2010. p. 1069–1072. (Cited on pages 33, 36, 37, 54 and 183.)

[110] Patenaude B, Smith SM, Kennedy DN, Jenkinson M. A Bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage. 2011;56(3):907–922. (Cited on pages 33, 37 and 183.)

[111] Bagci U, Chen X, Udupa JK. Hierarchical scale-based multiobject recognition of 3-D anatomical structures. Medical Imaging, IEEE Transactions on. 2012;31(3):777–789. (Cited on pages 33, 37 and 183.)

[112] Bernard F, Gemmar P, Husch A, Hertel F. Improvements on the Feasibility of Active Shape Model-based Subthalamic Nucleus Segmentation. Biomedical Engineering/Biomedizinische Technik. 2012;. (Cited on pages 33, 35, 54, 55 and 183.)

[113] Adiva E, Izmantoko YS, Choi HK. Comparison of Active Contour and Active Shape Approaches for Corpus Callosum Segmentation. 2013;16(9):1018–1030. (Cited on pages 33, 35 and 183.)

[114] Duta N, Sonka M. Segmentation and interpretation of mr brain images. an improved active shape model. Medical Imaging, IEEE Transactions on. 1998;17(6):1049–1062. (Cited on page 34.)

[115] Terzopoulos D, Fleischer K. Deformable models. The visual computer. 1988;4(6):306–331. (Cited on pages 38 and 183.)

[116] He L, Peng Z, Everding B, Wang X, Han CY, Weiss KL, et al. A comparative study of deformable contour methods on medical image segmentation. Image and Vision Computing. 2008;26(2):141–163. (Cited on pages 38 and 39.)

[117] Staib LH, Duncan JS. Boundary finding with parametrically deformable models. IEEE transactions on pattern analysis and machine intelligence. 1992;14(11):1061–1075. (Cited on page 38.)

[118] Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models. International journal of computer vision. 1988;1(4):321–331. (Cited on pages 38 and 39.)

[119] McInerney T, Terzopoulos D. T-snakes: Topology adaptive snakes. Medical image analysis. 2000;4(2):73–91. (Cited on pages 38, 39 and 54.)

[120] Lee JD, Tseng Yx, Liu Lc, Huang CH. A 2-D Automatic Segmentation Scheme for Brainstem and Cerebellum Regions in Brain MR Imaging. In: Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on. vol. 4. IEEE; 2007. p. 270–274. (Cited on pages 38, 39 and 54.)

[121] Osher S, Sethian JA. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. Journal of computational physics. 1988;79(1):12–49. (Cited on pages 38 and 39.)

[122] Wang Y, Staib LH. Boundary finding with correspondence using statistical shape models. In: Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on. IEEE; 1998. p. 338–345. (Cited on pages 38, 40, 54 and 56.)

[123] Duncan JS, Papademetris X, Yang J, Jackowski M, Zeng X, Staib LH. Geometric strategies for neuroanatomic analysis from MRI. Neuroimage. 2004;23:S34–S45. (Cited on pages 38, 40, 54 and 55.)

[124] Bekes G, Máté E, Nyúl LG, Kuba A, Fidrich M. Geometrical model-based segmentation of the organs of sight on CT images. Medical physics. 2008;35(2):735–743. (Cited on pages 38, 40, 54, 163, 165 and 166.)

[125] Lee M, Cho W, Kim S, Park S, Kim JH. Segmentation of interest region in medical volume images using geometric deformable model. Computers in biology and medicine. 2012;42(5):523–537. (Cited on pages 38 and 40.)

[126] Spinks R, Magnotta VA, Andreasen NC, Albright KC, Ziebell S, Nopoulos P, et al. Manual and automated measurement of the whole thalamus and mediodorsal nucleus using magnetic resonance imaging. Neuroimage. 2002;17(2):631–642. (Cited on pages 41, 47, 48, 54, 56 and 185.)

[127] Akselrod-Ballin A, Galun M, Gomori MJ, Basri R, Brandt A. Atlas guided identification of brain structures by combining 3D segmentation and SVM classification. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006. Springer; 2006. p. 209–216. (Cited on pages 41, 49 and 185.)

[128] Moghaddam MJ, Soltanian-Zadeh H. Automatic segmentation of brain structures using geometric moment invariants and artificial neural networks. In: Information Processing in Medical Imaging. Springer; 2009. p. 326–337. (Cited on pages 41, 47, 48, 54, 56, 85 and 185.)

[129] Zhou J, Chan K, Chong V, Krishnan S. Extraction of brain tumor from MR images using one-class support vector machine. In: Engineering in Medicine

and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the. IEEE; 2006. p. 6411–6414. (Cited on pages 41, 49 and 185.)

[130] Bauer S, Nolte LP, Reyes M. Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011. Springer; 2011. p. 354–361. (Cited on pages 41, 49 and 185.)

[131] Gasmi K, Kharrat A, Messaoud MB, Abid M. Automated segmentation of brain tumor using optimal texture features and support vector machine classifier. In: Image Analysis and Recognition. Springer; 2012. p. 230–239. (Cited on pages 41, 49 and 185.)

[132] Glotsos D, Tohka J, Ravazoula P, Cavouras D, Nikiforidis G. Automated diagnosis of brain tumours astrocytomas using probabilistic neural network clustering and support vector machines. International journal of neural systems. 2005;15(01n02):1–11. (Cited on pages 41 and 185.)

[133] Anbeek P, Išgum I, van Kooij BJ, Mol CP, Kersbergen KJ, Groenendaal F, et al. Automatic segmentation of eight tissue classes in neonatal brain MRI. PloS one. 2013;8(12):e81895. (Cited on pages 41 and 46.)

[134] Murino L, Granata D, Carfora MF, Selvan SE, Alfano B, Amato U, et al. Evaluation of supervised methods for the classification of major tissues and subcortical structures in multispectral brain magnetic resonance images. Computerized Medical Imaging and Graphics. 2014;38(5):337–347. (Cited on pages 41 and 46.)

[135] Larobina M, Murino L, Cervo A, Alfano B. Self-Trained Supervised Segmentation of Subcortical Brain Structures Using Multispectral Magnetic Resonance Images. BioMed research international. 2015;2015. (Cited on pages 41 and 46.)

[136] Cagnoni S, Coppini G, Rucci M, Caramella D, Valli G. Neural network segmentation of magnetic resonance spin echo images of the brain. Journal of biomedical engineering. 1993;15(5):355–362. (Cited on page 42.)

[137] Clarke L, Velthuizen R, Camacho M, Heine J, Vaidyanathan M, Hall L, et al. MRI segmentation: methods and applications. Magnetic resonance imaging. 1995;13(3):343–368. (Cited on page 42.)

[138] Kim EY, Johnson H. Multi-structure segmentation of multi-modal brain images using artificial neural networks. In: SPIE Medical Imaging. International Society for Optics and Photonics; 2010. p. 76234B–76234B. (Cited on pages 44 and 85.)

[139] Bengio Y, LeCun Y, et al. Scaling learning algorithms towards AI. Large-scale kernel machines. 2007;34(5). (Cited on page 45.)

[140] Webb AR. Statistical pattern recognition. John Wiley & Sons; 2003. (Cited on page 46.)

[141] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273–297. (Cited on pages 48, 65, 186, 205 and 209.)

[142] Vapnik VN, Vapnik V. Statistical learning theory. vol. 1. Wiley New York; 1998. (Cited on pages 48 and 205.)

[143] Burges CJ. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery. 1998;2(2):121–167. (Cited on pages 49, 65, 185, 186 and 212.)

[144] Honavar V. Artificial intelligence: An overview; 2006. (Cited on page 59.)

[145] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review. 1958;65(6):386. (Cited on page 65.)

[146] Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. DTIC Document; 1985. (Cited on page 65.)

[147] Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilistic language model. The Journal of Machine Learning Research. 2003;3:1137–1155. (Cited on pages 65 and 186.)

[148] Scholkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press; 2001. (Cited on pages 65, 87, 186 and 212.)

[149] Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. Advances in neural information processing systems. 2007;19:153. (Cited on pages 66, 77 and 187.)

[150] Cybenko G. Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems. 1989;2(4):303–314. (Cited on page 66.)

[151] Hinton G, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural computation. 2006;18(7):1527–1554. (Cited on pages 66 and 187.)

[152] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–507. (Cited on pages 66 and 187.)

[153] Bengio Y. Deep learning of representations for unsupervised and transfer learning. Unsupervised and Transfer Learning Challenges in Machine Learning, Volume 7. 2012;p. 19. (Cited on pages 66 and 187.)

[154] Schwarz G, et al. Estimating the dimension of a model. The annals of statistics. 1978;6(2):461–464. (Cited on page 66.)

[155] Erhan D, Bengio Y, Courville A, Manzagol PA, Vincent P, Bengio S. Why does unsupervised pre-training help deep learning? The Journal of Machine Learning Research. 2010;11:625–660. (Cited on page 67.)

[156] Palm RB. Prediction as a candidate for learning deep hierarchical models of data. Technical University of Denmark, Palm. 2012;25. (Cited on pages 67 and 101.)

[157] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105. (Cited on page 67.)

[158] Le QV. Building high-level features using large scale unsupervised learning. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE; 2013. p. 8595–8598. (Cited on page 67.)

[159] Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research. 2010;11:3371–3408. (Cited on pages 71, 76, 77, 91, 92, 126, 188 and 215.)

[160] Larochelle H, Erhan D, Vincent P. Deep learning using robust interdependent codes. In: International Conference on Artificial Intelligence and Statistics; 2009. p. 312–319. (Cited on page 73.)

[161] Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. ACM; 2008. p. 1096–1103. (Cited on pages 74, 187 and 188.)

[162] Maillet F, Eck D, Desjardins G, Lamere P, et al. Steerable Playlist Generation by Learning Song Similarity from Radio Station Playlists. In: ISMIR; 2009. p. 345–350. (Cited on pages 74 and 187.)

[163] Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research. 2010;11:3371–3408. (Cited on pages 74 and 187.)

[164] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11); 2011. p. 513–520. (Cited on pages 74 and 187.)

[165] Vincent P. A connection between score matching and denoising autoencoders. Neural computation. 2011;23(7):1661–1674. (Cited on pages 74 and 187.)

[166] Mesnil G, Dauphin Y, Glorot X, Rifai S, Bengio Y, Goodfellow IJ, et al. Unsupervised and Transfer Learning Challenge: a Deep Learning Approach. ICML Unsupervised and Transfer Learning. 2012;27:97–110. (Cited on pages 74 and 187.)

[167] Ng A. CS229 Lecture notes. CS229 Lecture notes. 2000;1(1):1–3. (Cited on page 78.)

[168] Bai W, Shi W, Ledig C, Rueckert D. Multi-atlas segmentation with augmented features for cardiac MR images. Medical image analysis. 2015;19(1):98–109. (Cited on pages 79 and 122.)

[169] Calonder M, Lepetit V, Ozuysal M, Trzcinski T, Strecha C, Fua P. BRIEF: Computing a local binary descriptor very fast. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2012;34(7):1281–1298. (Cited on page 80.)

[170] Kassner A, Thornhill R. Texture analysis: a review of neurologic MR imaging applications. American Journal of Neuroradiology. 2010;31(5):809–816. (Cited on page 80.)

[171] Aggarwal N, Agrawal R. First and second order statistics features for classification of magnetic resonance brain images. 2012;. (Cited on page 80.)

[172] Qurat-Ul-Ain GL, Kazmi SB, Jaffar MA, Mirza AM. Classification and segmentation of brain tumor using texture analysis. Recent Advances In Artificial Intelligence, Knowledge Engineering And Data Bases. 2010;p. 147–155. (Cited on page 80.)

[173] Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1989;11(7):674–693. (Cited on page 82.)

[174] Jin Y, Angelini E, Laine A. Wavelets in medical image processing: denoising, segmentation, and registration. In: Handbook of biomedical image analysis. Springer; 2005. p. 305–358. (Cited on page 82.)

[175] John P. Brain tumor classification using wavelet and texture based neural network. Int J Sci Eng Research. 2012;3(10). (Cited on page 82.)

[176] Criminisi A, Sharp T, Blake A. Geos: Geodesic image segmentation. In: Computer Vision–ECCV 2008. Springer; 2008. p. 99–112. (Cited on page 82.)

[177] Montagne C, Kodewitz A, Vigneron V, Giraud V, Lelandais S, et al. 3D Local Binary Pattern for PET image classification by SVM, Application to early Alzheimer disease diagnosis. In: Proc. of the 6th International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2013); 2013. p. 145–150. (Cited on page 83.)

[178] Pietikäinen M, Ojala T, Xu Z. Rotation-invariant texture classification using feature distributions. Pattern Recognition. 2000;33(1):43–52. (Cited on page 83.)

[179] Johnson HJ, McCormick M, Ibáñez L, Consortium TIS. The ITK Software Guide; 2013. *In press.* Available from: `http://www.itk.org/ItkSoftwareGuide.pdf`. (Cited on page 85.)

[180] Sarle WS, et al. Neural network FAQ. Periodic posting to the Usenet newsgroup comp ai neural-nets. 1997;. (Cited on page 87.)

[181] Hsu CW, Chang CC, Lin CJ, et al.. A practical guide to support vector classification; 2003. (Cited on page 87.)

[182] Breiman L, Spector P. Submodel selection and evaluation in regression. The X-random case. International statistical review/revue internationale de Statistique. 1992;p. 291–319. (Cited on page 89.)

[183] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011;2:27:1–27:27. (Cited on page 101.)

[184] Jannin P, Grova C, Maurer Jr CR. Model for defining and reporting reference-based validation protocols in medical image processing. International Journal of Computer Assisted Radiology and Surgery. 2006;1(2):63–73. (Cited on page 102.)

[185] Biancardi AM, Jirapatnakul AC, Reeves AP. A comparison of ground truth estimation methods. International journal of computer assisted radiology and surgery. 2010;5(3):295–305. (Cited on pages 108 and 168.)

[186] Fenster A, Chiu B. Evaluation of segmentation algorithms for medical imaging. In: Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the. IEEE; 2005. p. 7186–7189. (Cited on pages 110, 112 and 190.)

[187] Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945;26(3):297–302. (Cited on pages 110 and 111.)

[188] Andrews JR. Benefit, risk, and optimization by ROC analysis in cancer radiotherapy. International Journal of Radiation Oncology* Biology* Physics. 1985;11(8):1557–1562. (Cited on page 112.)

[189] Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1993;15(9):850–863. (Cited on pages 113 and 114.)

[190] Babalola KO, Patenaude B, Aljabar P, Schnabel J, Kennedy D, Crum W, et al. An evaluation of four automatic methods of segmenting the subcortical structures in the brain. Neuroimage. 2009;47(4):1435–1447. (Cited on pages 163, 165 and 166.)

[191] Fritscher KD, Peroni M, Zaffino P, Spadea MF, Schubert R, Sharp G. Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. Medical Physics. 2014;41(5):–. Available from: http://scitation.aip.org/content/aapm/journal/medphys/41/5/10.1118/1.4871623. (Cited on pages 165 and 166.)

[192] Duc AKH, Eminowicz G, Mendes R, Wong SL, McClelland J, Modat M, et al. Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. Medical physics. 2015;42(9):5027–5034. (Cited on pages 165 and 166.)

[193] Harrigan RL, Panda S, Asman AJ, Nelson KM, Chaganti S, DeLisi MP, et al. Robust optic nerve segmentation on clinically acquired computed tomography. Journal of Medical Imaging. 2014;1(3):034006–034006. (Cited on pages 165 and 166.)

[194] Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Mathematical programming. 1989;45(1-3):503–528. (Cited on page 203.)

[195] Karush W. Minima of functions of several variables with inequalities as side constraints. Master' s thesis, Dept. of Mathematics, Univ. of Chicago; 1939. (Cited on page 208.)

[196] Kuhn H, Tucker A. pp. 481–492 in: Nonlinear Programming. In: Proc. 2nd Berkeley Symp. Math. Stat. Prob.(J. Neyman, ed.), Univ. of Calif. Press, Berkeley, CA. vol. 14; 1951. . (Cited on page 208.)

[197] Fletcher R. Practical methods of optimization. John Wiley & Sons; 2013. (Cited on page 209.)

[198] Abdi MJ, Hosseini SM, Rezghi M. A novel weighted support vector machine based on particle swarm optimization for gene selection and tumor classification. Computational and mathematical methods in medicine. 2012;2012. (Cited on page 213.)