CHALMERS
UNIVERSITY OF TECHNOLOGY

Tissue Segmentation of Head and Neck for Hyperthermia Treatment Planning

Master's thesis in Engineering Mathematics

ERIKA EK

Department of Electrical Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2018

Master's thesis 2018:30

Tissue Segmentation of Head and Neck for Hyperthermia Treatment Planning

Erika Ek

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Division of Signal Processing and Biomedical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2018

Tissue Segmentation of Head and Neck for Hyperthermia Treatment Planning
ERIKA EK

Supervisor: Professor Fredrik Kahl, Department of Electrical Engineering
Examiner: Assistant Professor Hana Dobšíček Trefná, Department of Electrical Engineering

Master's Thesis 2017:30
Department of Electrical Engineering
Division of Signal processing and Biomedical engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Segmentation of the mandible (magenta), brainstem (blue) and the parotid glands (green) obtained through the means of deep learning. The images represent three orthogonal views in the median plane of the organ structures and they are constructed by the use of data from [1]

Typeset in LaTeX

Tissue Segmentation of Head and Neck for Hyperthermia Treatment Planning
ERIKA EK
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Head and neck cancer is the ninth most common malignancy in the world with a survival rate of around 40-50 %. One promising treatment method is local hyperthermia applied using an antenna array. The antennas are set to radiate microwaves of a certain amplitude and phase selected for interference in the tumor. These settings are derived in a treatment planning process which is highly dependent on an accurate tissue segmentation of a 3D image representation of the patient, which commonly is CT based.
The aim of this thesis is to present an automated 3D segmentation method for CT images of the head and neck region through deep learning. Its performance will thereafter be compared to the current state-of the art. This is achieved using a further developed version of the U-net convolutional neural network for 3D segmentation. The network is trained using manually segmented CT data sets of the head and neck region to segment three different organs: the mandible, brainstem and the parotid glands. The data sets were obtained from the MICCAI 2015 segmentation challenge and the final results are compared to those obtained by the top three teams from this competition.
An ablation study was used to compare different data processing and augmentation types. The highest performing model was obtained when training the network on the whole image using a combination of reflection, distortion and rotation augmentation. Generally, the implemented segmentation method performed similarly as the teams from the MICCAI challenge according to the Dice coefficient. However, the method outperformed the teams when considering the Hausdorff distance and the contour mean distance for all organs. The conclusion is therefore that it could be advantageous to use this deep learning segmentation method for tissue segmentation of organs of interest for hyperthermia treatment.

Keywords: Tissue segmentation, deep learning, supervised learning, convolutional neural networks, data augmentation, medical imaging, computed tomography, head and neck cancer, hyperthermia

Acknowledgements

I would first of all like to thank my supervisor Fredrik Kahl and my examiner Hana Dobšíček Trefná for the opportunity of being involved in such an interesting project and for their support throughout the process. Furthermore, I'm especially grateful for all the assistance I've received from people of the department which have had no official obligation to help me but have done so solely out of kindness and generosity. This especially true for Samuel Scheidegger, Jennifer Alvén, Måns Larsson and Massimiliano Zanoli. Finally, I would of course also like to thank my closest friends and family for a constant moral support and for lifting me up when I was down.

Erika Ek, Gothenburg, June 12, 2018

Contents

List of Figures

List of Tables

1
Introduction

Head and neck cancer (HNC) represents the world's ninth most common malignancy with a survival rate of around 40-50 % [2], [3]. There are several possible causes of HNC but amongst all etiological factors, excessive alcohol consumption and tobacco use are considered to be the most prominent [4]. Accounts for HNC are expected to increase worldwide for both genders, also targeting the younger population more prominently. Both the current and estimated future burden of these types of cancer are observed to shift more and more towards less developed countries. This can be highly problematic due to the fact that these may be ill equipped to deal with such an increase [2].

The use of hyperthermia (HT), especially applied locally, is considered to be promising in treatment of HNC [5]. An accurate 3D segmentation of the patient is a crucial part of treatment planning of microwave HT. This type of HT is applied using an antenna array and the settings of these are directly determined from this planning step. Consequently, a more accurate model of the patient will provide more suitable settings. In addition, the decision of whether or not to treat the patient in the first place is based on the results of this process [6]. Currently, head and neck models are generated by manual tissue delineation in a set of computed tomography (CT) images. This method is both labor intensive and time consuming for the operator, where a manually segmented 3D model can take around 5-6 hours [6],[7]. Furthermore, both inter-observer and intra-observer variability affect the reproducibility of such models [6]. The demand for an automated 3D segmentation method to produce anatomically accurate models is therefore high.

Most of the different tissues in the head and neck region possess highly varying di-electrical and thermal properties. It is therefore of great importance to segment these to to be able to construct an accurate treatment plan. There is one already existing automated segmentation method constructed for this purpose which uses a combination of multiatlas and intensity modelling in a graph cut framework. This approach, in difference to manual segmentation, takes around 3 hours/patient [6],[7]. In other cases, a deep learning approach has proven to be superior to a corresponding multiatlas solution [8]. The goal of this project is to construct an alternative segmentation method using deep learning and discuss the differences in performance.

1.1 Aim

The aim of this project is to construct an automated 3D segmentation method through the means of deep learning for a set of CT images representing the head and neck region. The final purpose of the resulting program is to serve as a step in the treatment planning process for antenna based local hyperthermia for patients suffering from head and neck cancer.

1.2 Demarcations

The organs of interest for hyperthermia purposes in the head and neck region are the cerebrum, cerebellum, brainstem, spinal cord, optical nerve, sclera, cornea, eye vitreous humor, lens, cartilage and thyroid, illustrated in Figure 1.1. To create a full model of the patient, the currently existing automatic segmentation method for this purpose uses thresholding of the of the CT voxel values to segment muscle, bone and the lungs [7].

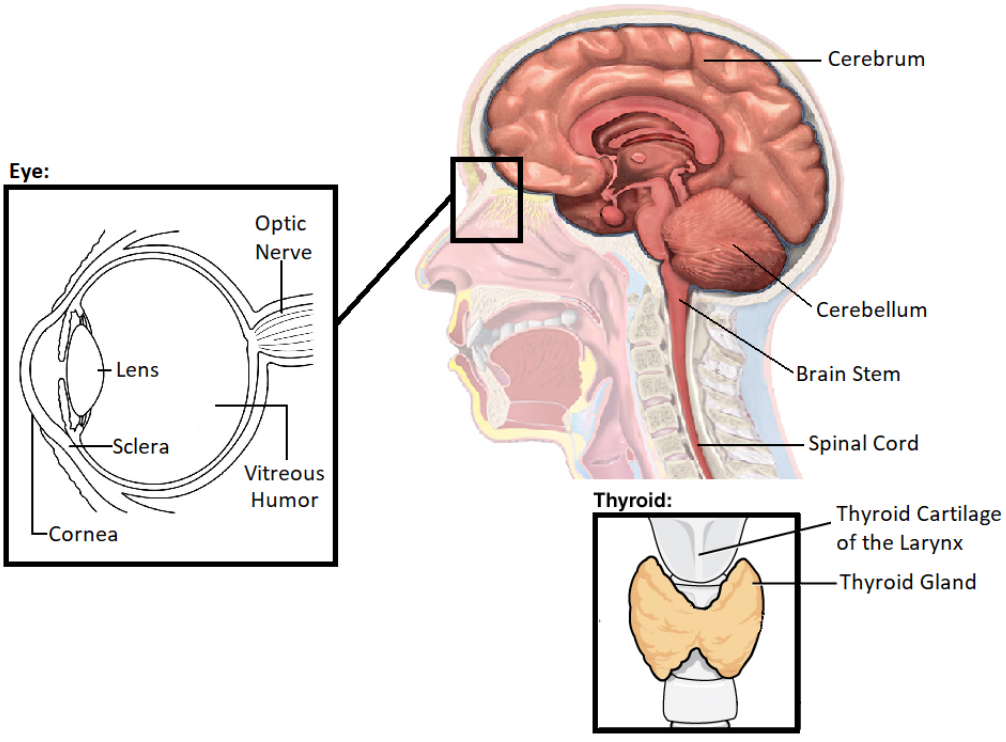


Figure 1.1: This image illustrates the organs of interest for segmentation in hyperthermia treatment planning. The image was constructed based on [9], [10] and [11].

Due to a highly time consuming legal process, ground truth data used by the current state-of-the-art was not made available in time for this project. In its place, segmented CT images of the head and neck region from the MICCAI auto-segmentation challenge 2015 were used. The target organs of this set are: the brainstem, mandible,

left and right optic nerves, optic chiasma, left and right parotid glands and the left and right submandibular glands [1]. These are illustrated in Figure 1.2, apart from the brainstem which can be seen in Figure 1.1. Out of these organs, three out of six structures were selected for segmentation given the limited amount of time for this project. The three organs were chosen based on which organs the participants of the MICCAI 2015 challenge scored the highest on, namely the mandible, the brainstem and the parotid glands in that same consecutive order.

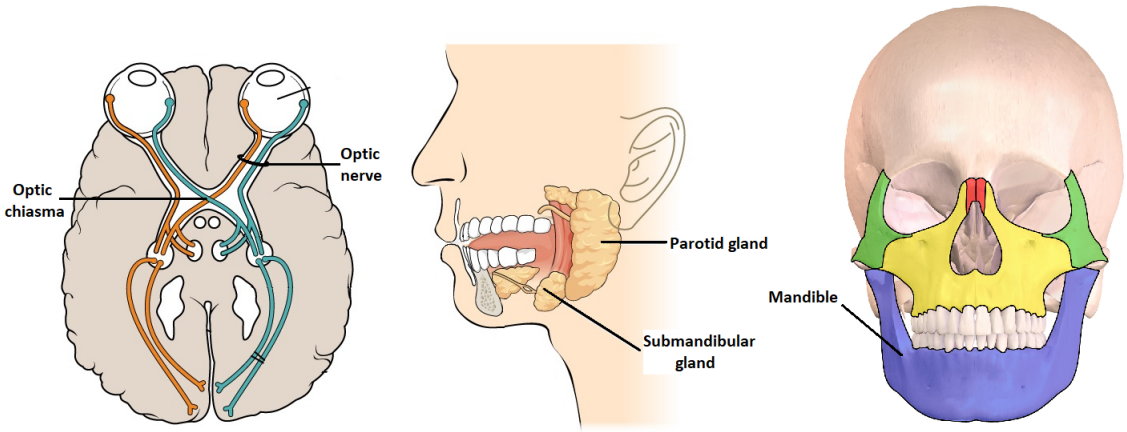


Figure 1.2: This image illustrates the organs of interest for segmentation during the MICCAI 2015 segmentation challenge. The image is based on [12], [13] and [14].

Some submitted and published segmentation approaches to this challenge were atlas based, as is the current segmentation approach for hyperthermia purposes [1]. The comparison in performance between atlas based segmentation and segmentation through deep learning can thus still be made. Remaining submitted approaches were model based, and consequently this deep learning approach can be put up against two common approaches within the medical segmentation field [1].
The data used in the MICCAI 2015 challange is not fully annotated, i.e. not all the organs are segmented in all data sets. For this reason the network is trained separately for each organ to be able to use the maximum amount of available data for each organ.
Computed tomography (CT) images are highly relevant to use as basis for radiotherapy treatment planning due to the fact that they provide the proton density of the tissues and they are considered to be more geometrically accurate than for instance magnetic resonance imaging (MRI). However, MR images are known to have a higher soft tissue contrast which could be of aid to an automatic segmentation process [6]. This project will solely focus on segmentation of CT, since this is the format of the available ground truth data.
Automatic image segmentation can be achieved by several different means. This project will investigate the performance of a a deep learning approach, more specifically a convolutional neural network (CNN), as the only implemented method. The results will be compared to those obtained by the top three participating teams of the MICCAI 2015 challenge [1].
There are several ways of evaluating the quality of an image segmentation com-

pared to the ground truth [15]. In this project, the Dice similarity coefficient (DSC), the mean surface distance (MSD) and the Hausdorff Surface Distance (HSD) will be used. These quality indicators are considered suitable since they are selected for evaluating the submitted segmentation approaches in the MICCAI 2015 challenge[1]. These same quality indicators are also used by the current state-of-the-art in hyperthermia segmentation, which increases the comparability factor [7].

1.3 Specification of Issue Under Investigation

How does a deep learning approach compare to corresponding atlas and/or model based approaches for 3D segmentation of CT images of the head and neck region? Based on these results, would it be advantageous to implement such a method as an alternative to the currently used segmentation tool used in hyperthermia treatment planning?

1.4 Related Work

Apart from machine learning approaches to automatic image segmentation, there are two other main categories which are especially common within the field of medical image segmentation: model based segmentation and atlas based segmentation [16]

1.4.1 Model Based Segmentation

One basic approach to automated medical image segmentation is based on the assumption that organs are of a repetitive geometrical shape. By this assumption, they can be modelled probabilistically for variation in geometry and shape. Using this as a constraint, the image can be segmented by an initial registration of the training data, a probabilistic representation of the variation of this data and finally a statistical relation between model and image [16]. Two examples of such approaches are, the most common approach, the active shape model (ASM) and the active appearance model (AAM), which is considered to be the most powerful [17], [18].
Model based approaches to image segmentation are known to be highly robust when dealing with artifacts and noise in the image data [17]. They do, however, require manual interaction in placing an initial model with corresponding suitable parameter initialization [16].

1.4.2 Atlas Based Segmentation

Multi-atlas segmentation is one of the most successful and widely used segmentation methods for biomedical imaging [19]. These methods rely on the existence of reference images, called atlases, which are images combined with a manual segmentation of the target area. One atlas can thereafter be used to segment a new patient image by the use of image registration [20]. By having multiple atlases it is possible to register them all to the patient image and derive the new segmentation by performing

a majority voting for each voxel. The final segmentation will hence consist of the voxels which are voted for by a majority of the atlases [21].
One comparative advantage to this method is its ability to segment an image where there is no clear relation between the region and the pixel intensity [20], [22]. This method is, however, quite computationally expensive due to the multiple image registration steps [19], [22]. Atlas guided approaches are also generally better suited for segmentation of structures which are stable over the population of the study [23].

1.4.3 Segmentation through Machine Learning

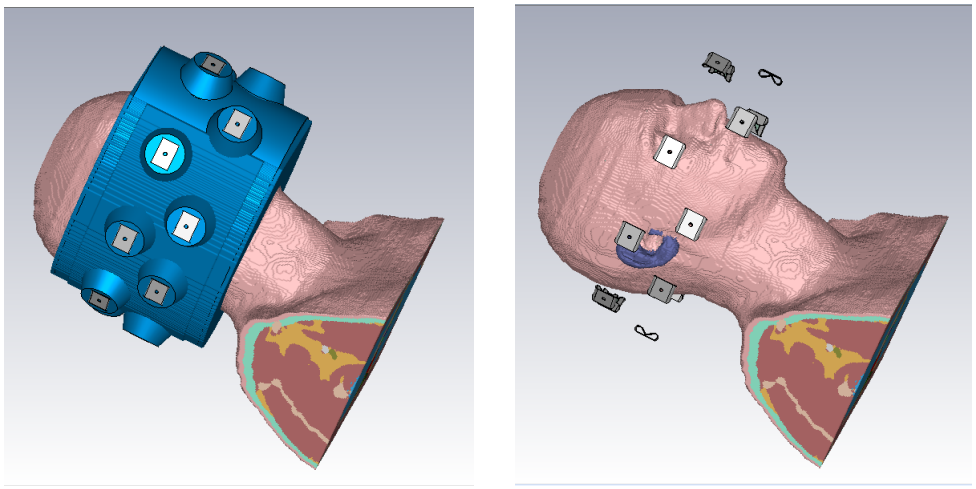
There are two main categories of this segmentation approach: supervised and unsupervised learning. During supervised learning is common to use artificial neural networks which are trained using manually segmented images as ground truth. Such networks are composed of interconnected elements, often denoted neurons, whose connection is denoted weights. These are adapted during an optimization process to solve a specific task. One main advantage to this approach is its ability to learn adaptively according to the training data and to self-organize depending on the input data. However, the performance of such approaches are both negatively affected by the presence of noise as well as sensitive to the choice of training parameters [16].
Unsupervised methods are not dependent on any type of ground truth data. This is commonly achieved by the use of clustering, the process of finding natural grouping clusters in a multidimensional feature space. There are several different algorithms for this purpose, none of which is known to be superior for a particular application. This type of segmentation has the advantage of minimal operator interaction and reproducibility of the results. However, such techniques may not result in the optimal solution and there is a need for operator intervention for error correction in the case the resulting segmentation is found inadequate [16].
In this project a supervised learning approach to image segmentation will be used with a deep artificial neural network. The main contribution of this work will be to first and foremost evaluate the performance of such an approach on certain organs of the head and neck region and to compare its outcome to that of other approaches using the same data. No other corresponding deep learning approach to segment these particular organs could be found in literature. Secondly, this comparison will serve as a basis for the decision of whether or not it is deemed beneficial to use a deep learning approach to the segmentation step in hyperthermia treatment planning.

2
Background

In this chapter, the concept of hyperthermia treatment will be further explained along with an introduction to medical imaging techniques focusing on computed tomography imaging. Thereafter follows a general introduction to image registration and segmentation. Finally, a more detailed explanation of the inner workings of a deep learning approach using convolutional neural networks will be presented.

2.1 Hyperthermia Treatment of Head and Neck Cancer

As previously stated, hyperthermia (HT) is considered to be a promising treatment option of head and neck cancer (HNC), together with immunotherapy, radiotherapy, chemotherapy and surgery. It is commonly used in combination with surgery, radiotherapy and chemotherapy, especially to reduce the toxicity of the latter two [5]. Local HT aims to selectively heat the tumor to a temperature of 40-44°C using a heating device. The goal is to elevate the temperature of the tumor to 43°C without harming thermo-sensitive tissue nearby [6]. The elevated temperature has been proven to enhance the effect of radiation, enhance cytostatic drugs and cause heat-induced cell death [24].
Generally, the process of local HT treatment is to use a contacting medium in between the skin and the antennas or applicators placed to heat the tumor tissue using electromagnetic waves or ultrasound [5]. There are several applicator types which have been used in clinic, which can be divided into three main categories: ultrasound therapy, microwave therapy and near-infrared photothermal therapy [5]. This project will address the treatment planning of microwave therapy applied using an array of antennas. A model of such an applicator for a tumor in the head and neck region is presented in Figure 2.1.



(a) **(b)**

Figure 2.1: This figure present models of an antenna array setup to be used for hyperthermia treatment of head and neck cancer. In Subfigure 2.1a, the blue material is a cooling system called bolus, in Subfigure 2.1b this structure is excluded to more clearly visualize the patient and the antennas. The images are constructed with the modelling program CST Studio Suite.

In microwave therapy, the antennas can be set to radiate at different frequencies, usually within the range of 430-2450 MHz, depending mainly on the tumor's size and location [5]. Before applying local HT treatment to a patient, it must be planned so that each antenna within the array will radiate at a specific phase and amplitude yielding optimal heating for a particular tumor [7]. An overview of the different steps of this type of HT treatment planning is presented in Figure 2.2.

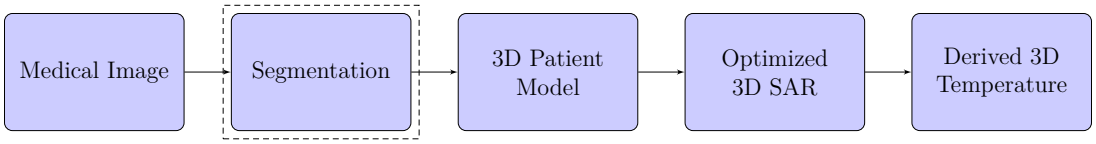


Figure 2.2: This figure presents the different steps of the hyperthermia treatment planning procedure for using microwaves radiating from an antenna array to locally treat head and neck cancer. The marked box represents the step in which this project will focus on.

The final step of the treatment planning process, the optimization, is often performed so that the absorbed power, called specific absorption rate (SAR), will be as high as possible in the target tissue whilst sparing the normal and thermo-sensitive ones from excessive heating [7]. This project will focus on the one particular part of och the treatment planning process which is marked in Figure 2.2: the segmentation of a medical image representation of the patient.

2.2 Medical Imaging Techniques

There are several different ways of depicting the inside of the human body and these can be divided into two main types: invasive and noninvasive techniques. Invasive techniques include methods such as surgery and endoscopy involving cutting the body open and/or inserting an object into the body, which expose the patient to a risk of both trauma and damage [25]. Noninvasive alternatives to these include for instance computed tomography (CT) and magnetic resonance imaging (MRI), both of which are widely used for different segmentation techniques [25],[26].

2.2.1 Computed Tomography (CT)

Computed tomography (CT) is an imaging technique which uses several shots of projection tomography images, commonly referred to as as x-ray images, to depict a cross section of the body [25]. The technique of x-ray imaging is based on the transmission of ionizing radiation through the body. Different body tissues attenuate the beam of ionizing radiation separately depending on their composition, which consequently means that a homogeneous beam radiated at a body will leave a exiting beam which will serve as a shadow of its interior structure. The intensity level of the detected exiting radiation will thus provide anatomical information based on the tissues' attenuation abilities [25]. Dense tissues, such as bone, absorb much radiation depicting it as white whereas soft tissues will absorb less and will be presented in gray. Cavities filled with air absorb little, and are thus almost black [16]. CT images are taken using a device similar to the one depicted in Figure 2.3.

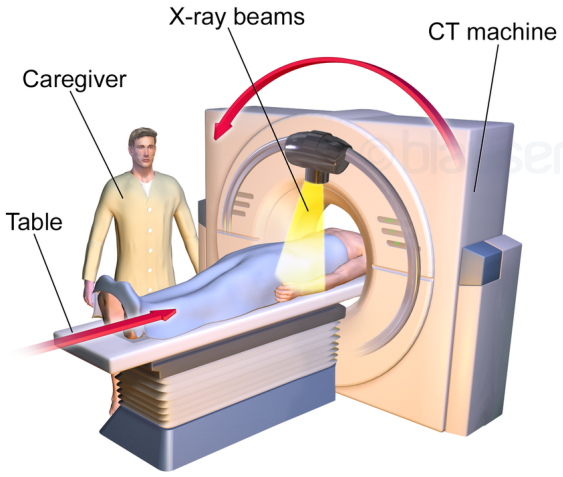


Figure 2.3: This is an illustration of a machine used to create computed tomography (CT) representations of patients. The image is a slightly modified version of an image from [27].

Distinguishing CT from simple x-rays is the illustration of cross sections of the body instead of simply showing projections. This removes the overlaying structures of organs with the cost of lower resolution and a higher radiation dose per image. When these cross section images are stacked together they have the advantage of

providing a 3D-model of the patient and thus enabling other angles in which the target area can be visualized [25]. This is illustrated in Figure 2.4, where a head and neck CT scan is presented in three orthogonal planes in the center of the model.

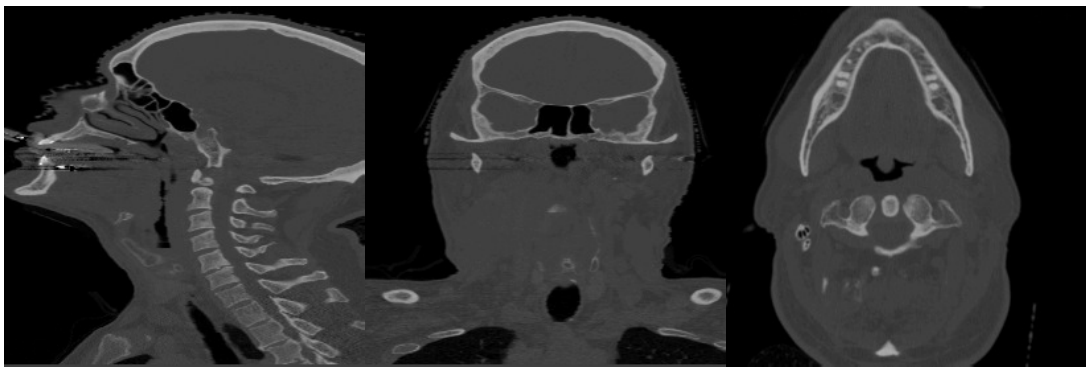


Figure 2.4: This figure presents a CT scan of the head and neck region from 3 orthogonal views. The data used to construct this image is from [1]. Note that this image is slightly compressed in the z-dimension.

One disadvantage of CT scans is that exposure to such radiation, even at low doses, increases the risk for cancer. At high doses, it can even cause burns or cataract formation [25]. Another issue is the difficulty to distinguish between different soft tissues. These have similar composition and thus similar attenuation properties, but in some cases it can be necessary to clearly distinguish between such tissues, as for instance in brain imaging. Both of these issues are nonexistent for MR imaging [25],[16].

2.3 Image Registration

Image registration is the process of geometrically aligning two or more images with the same motif. The majority of registration methods involve detecting image features, matching these to estimate a suitable image transformation and finally transforming the image. All steps, except the final transformation, can be executed in many fashions. This is especially true for the estimation of the image transformation, where both optimization method and transformation method is varied depending on the task at hand [28]. The general description of image transformation is:

$$\hat{\mathbf{x}} = \mathcal{T}(\mathbf{x}), \tag{2.1}$$

where the point $\mathbf{x}$ is transformed by $\mathcal{T}$ to a transformed point $\hat{\mathbf{x}}$. The transformation $\mathcal{T}$ is an operator which is defined depending on the sought degrees of freedom and of course the image dimension [29].

2.3.1 Image transformation

There are several different means of transforming an image, some of which are rigid, non-rigid and affine transformation. In a rigid transformation, all distances are preserved and the image is assumed to only have been rotated and translated. The transform can thus be described as:

$$\hat{\mathbf{x}} = R\mathbf{x} + t, \tag{2.2}$$

where R is a rotation matrix [29]. In this case with three dimensions, this rotation matrix is defined as:

$$R = \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 \\ \sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y \\ 0 & 1 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix}, \tag{2.3}$$

where θ_x, θ_y and θ_z are the rotation angle in each direction respectively [29]. The non-rigid transformation, also known as the similarity transformation, has a definition which is only slightly different from the rigid transformation in Equation (2.2). The difference is that it also allows image scaling multiplied with the rotation matrix of a chosen factor [21].
The affine transformation is even less restrictive since it also allows anisotropic stretching of the image. It can thus be described as:

$$\hat{\mathbf{x}} = A\mathbf{x} + t, \tag{2.4}$$

with no further restrictions on the elements a_{ij} of the matrix A [29]. In the context of medical imaging, this can be used to represent differences in, for instance, bone structure.

2.4 Image Segmentation

Imaging is a mean of transferring information and the task of understanding and extracting this information is considered to be an important field within digital image technology. Image segmentation is the first step in this process, the basis of image recognition and one of the most prominent areas of computer vision and image processing [30]. The main idea of image segmentation is to extract an area of interest from the image by grouping together elements which appear to be of the same nature. There are currently several different algorithms used for this purpose, some of which are: region based-, edge detection-, deep learning segmentation and segmentation based on clustering. This technology has several different application areas, one of which is medical image processing [30].

2.4.1 In the Context of Medical Imaging

The use of image modalities such as CT and MRI for both diagnostics and treatment planning is growing and so does also the demand for computer assistance in image

analysis for the radiologists [26]. Currently available techniques are specific to both application and modality due to differences in contrast and artifacts [16]. Basically, the image going through this process will be divided into segments having similar properties such as contrast, texture, brightness and intensity. As for medical images, these are often particularly complex, which makes it problematic to directly utilize methods within image processing. One reason for this is that the regions of interest usually differ somewhat in size and location between different patients [26].
Medical image segmentation is commonly divided into three categories: manual-, semi-automatic and fully automatic segmentation. Manual segmentation consists of a medical expert the drawing organ boundaries. As previously mentioned, this is a highly time consuming task and the variance in both inter- and intra-observability affects the reproducibility. The latter is highly problematic since such segmentation is used when implementing fully automatic segmentation algorithms. In spite its dependence on the rater it is widely used, especially in clinical trials where the time is less of a factor [26]. One example of manual segmentation of the mandible, brainstem and the parotid glands in the head and neck region is presented in Figure 2.5.

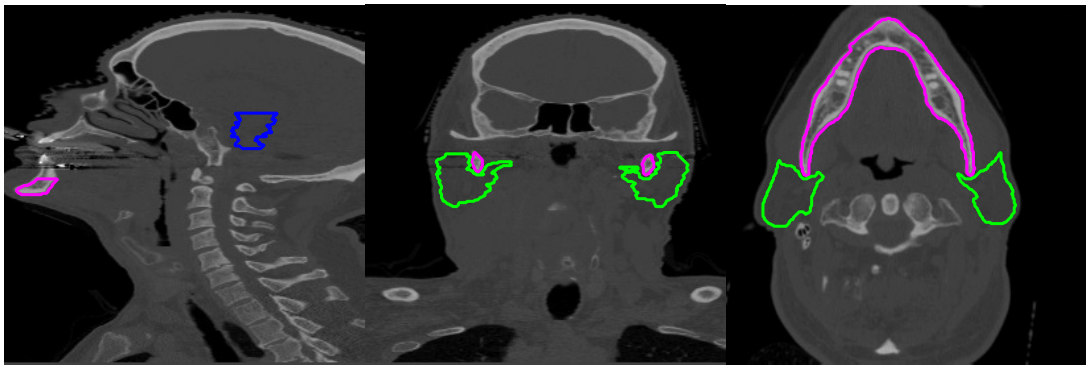


Figure 2.5: This figure presents a manual segmentation of the mandible, brainstem and the parotid glands in three orthogonal views i the middle of a head and neck CT model. The mandible is magenta colored and is visible in all three views, the brainstem is blue and is only visible in the left view and the parotid glands are green and visible in the two rightmost views. The image is constructed by the use of data from [1].

Semi-automatic segmentation algorithms require minimal human interaction to initialize the process or to correct segmentation results manually [26]. These can for instance include estimating the segmentation of some tissues using geometrical models, such as a sphere for eyeball or similar [6].
The fully automatic approach performs the segmentation without human interference. In such cases, human knowledge is incorporated during the implementation, as for model-based approaches with soft computing methods. Manually performed segmentation requires specialized knowledge along with high level visual processing and is thus a highly demanding task for an automated algorithm to perform. These

are therefore not widely used in clinical practice but they could be advantageous when dealing with large image sets [26].
In constructing a 3D model of a patient during hyperthermia treatment planning it would be highly advantageous to use a fully automatic segmentation approach to minimize the time factor of this step [6].

2.4.2 Evaluating Segmentation Accuracy

It is of great importance to be able assess the accuracy and quality level of a segmentation [15]. The most common quantitative standard measure in medical segmentation is the dice similarity coefficient (DSC) [26]. It is defined as follows:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2TP}{2TP + FP + FN} \tag{2.5}$$

where X, Y represent binary 3D images, $|\bullet|$ the number of voxels in the set, TP, FP and FN stand for true positive, false positive and false negative, respectively [26], [15]. This quantification is visualized in Figure 2.6.

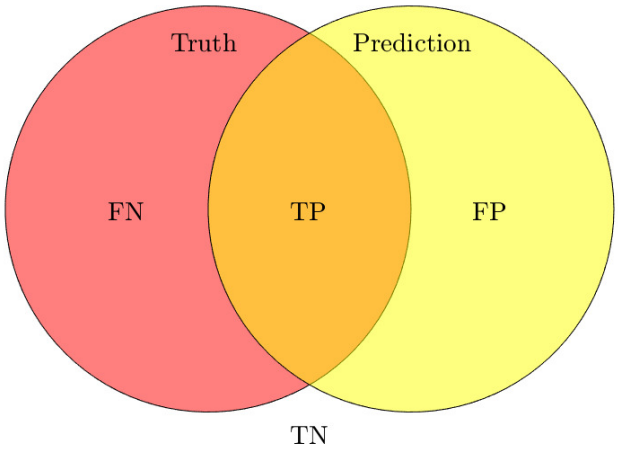


Figure 2.6: This figure presents a visualization of the dice similarity coefficient (DSC), described in Equation (2.5), where FN, TP, FP, and TN stand for false negative, true positive, false positive and true negative, respectively. *Truth* marks the true segmentation and *Prediction* the predicted segmentation.

Note that, from observing Equation (2.5), it is clear that a fully correct segmentation will result in $DSC = 1$ and a completely incorrect segmentation in $DSC = 0$.
Another way of quantifying ground truth compatibility is to calculate the contour mean distance (CM), also known as the mean surface distance (MSD). This measures the mean distance between the surface of the volumes and is defined below:

$$CM = \max\Big(m(X,Y), m(Y,X)\Big), \tag{2.6}$$

$$m(X,Y) = \underset{x\in X}{\text{mean}} \min_{y\in Y} ||x-y|| \tag{2.7}$$

where $x \in X$ and $y \in Y$ are contour points of the volumes X,Y and $||\bullet||$ denotes the Euclidean distance between x and y. This measures the average mismatch between the boundaries of X and Y, and should thus ideally be as low as possible [1]. Similarly, the maximum Hausdorff Surface Distance (HSD) is used to measure the maximum distance between points on the surface of volumes. This measure is defined as:

$$HD = \max\Big(h(X,Y), h(Y,X)\Big), \tag{2.8}$$

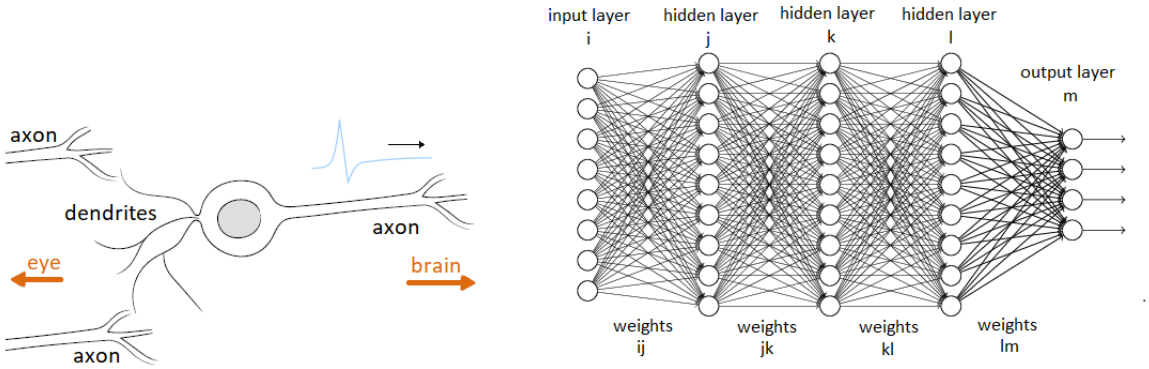
$$h(X,Y) = \max_{x\in X} \min_{y\in Y} ||x-y|| \tag{2.9}$$

with the same notation as in Equation (2.7). Another measure used is the 95 % HD, where the 95th percentile of the distance between boundary points in X and Y is considered. Doing so reduces the impact of outlier measurements on the evaluation of the overall segmentation quality [1]. This measure consequently quantifies the maximum mismatch between the boundaries of X and Y, and should also ideally be as low as possible.

2.5 Convolutional Neural Networks (CNN)

A convolutional neural network (CNN) is a specific type of neural network where a characteristic grid-like structure is used to process data. Its name stems from the usage of the mathematical convolution operation instead of a traditional matrix multiplication in at least one of the network layers [8]. CNN:s have had an important role in the context of deep learning as they were the first deep models to perform well. They are also among the first neural networks used in important commercial applications and they remain at the forefront of such solutions to this day. Their success is considered to be a key example of the potential advantage of applying neuroscientific insights to machine learning implementations [8].
These deep feed forward networks are designed to mimic the image processing steps carried out by neurons in the visual cortex. This system is highly complex and therefore difficult to model, especially since all elements of the mammalian vision system have not been completely described yet [8]. There is therefore an element of chance involved in creating such a model and the superiority of well functioning networks is not fully understood [8].
Physiologically speaking, an image is registered by light sensitive neurons in the eye and the information is subsequently transported through many steps of neural communication before being analyzed in the brain. As illustrated in Subfigure 2.7a, a neuron can receive information from several other neurons by a dendrite-connection to their axons. This information is thereafter processed and transferred to other

neurons through its own axon [21]. This manner of processing an image is mimicked using layers of models of such neurons, also called perceptrons, connected through weight factors, see Subfigure 2.7b [21], [31].



(a) **(b)**

Figure 2.7: Subfigure 2.7a shows a model of a neuron [21] and Subfigure 2.7b presents a schematic overview of a neural network [31].

When the image is run through the network, the output from the previous layer to the next is often referred to as a feature map [31]. Inter-neuron communication naturally depends on the seen image but exactly how is not fully understood which makes it complex to imitate [8].

2.5.1 Deep Networks

Applying a neural network to a large data set, as for instance medical 3D images, result in a fully connected network with a large amount of weights which would difficult to train [21]. Note that the network presented in Subfigure 2.7b only has at most 9 neurons but weight matrices with $9 \times 9 = 81$ elements in between the layers. In such cases, it is common to use a smaller set of weights applied to the image as a filter by convolution. This means that a set of filters are learned instead of weights for individual neurons [21].
Reducing the set of parameters by introducing weight-filters opens up the possibility of constructing and training much deeper networks. The structure of deep CNN:s can vary in for instance depth, branching, filter size and layering. There are, however, a number of well established techniques which are independent of the structure, three of which are max pooling, dropout and application of a rectified linear unit [21].
Max pooling is used to reduce the dimension of the target which is fed to the network. This can be done using different stride sizes depending on the wanted down sample rate. Basically, the image is divided into regions of the chosen stride size whereafter the maximim value is chosen as the new voxel representation. A simple 2D example is presented in Figure 2.8 [21].

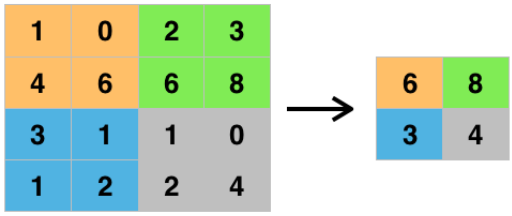


Figure 2.8: An illustration of max pooling of a 4×4 image with stride 2, from [32].

It is also common to use a method known as dropout when training a deep CNN. Dropout is employed to prevent a network from getting too adjusted to the training data, a principle known as overfitting. According to this method, a random portion of the neurons in selected layers are shut down in each training step, being excluded from training. This creates an illusion of a slightly different image and can thus increase the network's generalization capability [21].
After each convolutional layer it is common to introduce an activation layer by using a rectified linear unit (ReLu). This is an activation function defined as:

$$f(x) = max(0, x), \tag{2.10}$$

and it is applied to each voxel of the input image [33]. It will simply set negative input values to zero while preserving the positive values. Employing this function increases the amount zero-valued elements, which in turn reduces the computation cost and is also known to introduce a sparse representation which leads to certain mathematical advantages of the network [33].
The network is designed to have an output of a certain dimension depending on the sought output [21]. In the case of image segmentation, the output is a probability map which indicates the probability of a certain region belonging to a certain class [34].

2.5.2 Learning the Network Weights

One common approach is to randomly initialize the weights and update them according to the network response to the image [21]. The network parameters, weights, are in most cases optimized using a version of gradient descent. A chosen loss function of the network parameters is minimized by updating the parameters in the opposite direction of the gradient of this function. This, meaning that the weights are in each step updated using a chosen update rule, specifically defined depending on the chosen optimizer, where the new weights are derived as a function of the current weights and the negative gradient of a chosen loss function [35].
In each step, a set of training images is randomly selected to be run through the network, referring to the weights of each layer are applied to the images. This is necessary in order to calculate the gradient of the loss function in each layer. Thereafter, the weights can be updated through what is called a *backpropagation*, where starting at the output layer, the weights between each layer are updated step by step using the previously calculated gradients [21].

2.5.2.1 Stochastic gradient descent

Stochastic gradient descent (SGD) is an algorithm which is frequently used to train neural networks. In SGD the target parameters θ, in this case the weights, are updated by adding the negative gradient of the loss function L multiplied by a chosen learning rate μ [21]. The update rule can thus be defined as follows:

$$\theta_{t+1} = \theta_t - \mu \nabla L_t, \tag{2.11}$$

for a time step k to $k+1$ where the gradient ∇L_t is calculated with respect to θ_t and is a function of θ_t. To help the algorithm in converge whilst damping its oscillation, it is possible to add fraction γ of the update vector to the update rule [35]:

$$\theta_{t+1} = \theta_t - v_t, \tag{2.12}$$
$$v_t = \gamma v_{t-1} + \mu \nabla L_t. \tag{2.13}$$

The parameter γ is often referred to as the momentum term.

2.5.2.2 The Adam Optimizer

An alternative to SGD is the Adaptive Moment Estimation (Adam). This method computes individual adaptive learning rates for the parameters using estimates of the first and second moment of the gradients [36]. This method stores an exponentially decaying average of past squared gradients v_t whilst also keeping an exponentially decaying average of past gradients m_t, similar to momentum:

$$m_t = \beta_1 m_{(t-1)} + (1 - \beta_1) \nabla L_t, \tag{2.14}$$
$$v_t = \beta_2 v_{(t-1)} + (1 - \beta_2)(\nabla L_t)^2, \tag{2.15}$$

where β_1 and β_2 are decay rates, m_t and v_t are estimates of the first and second moment, the mean and uncentered variance, respectively [35]. To counteract an initial bias trend towards zero for m_t and v_t, bias-corrected first and second order estimates are computed as:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \tag{2.16}$$
$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}. \tag{2.17}$$

Finally, the weights are updated using the following update rule:

$$\theta_{t+1} = \theta_t - \frac{\mu}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t, \tag{2.18}$$

where μ again represents the learning rate and the parameter ϵ is set as a small constant value [35].

2.5.2.3 The Loss Function

A common choice of loss function is a cross-entropy loss using softmax, particularly so when there is a need to separate more than two classes. The neuron values x_k can be converted to class probabilities p_k using the softmax function [21]:

$$p_k = \frac{e^{x_k}}{\sum_l e^{x_l}} \quad \Rightarrow \quad \sum_k p_k = 1. \tag{2.19}$$

The cross-entropy loss of these class probabilities is defined as:

$$L(\theta) = -\sum_{i \in S_0} p_{0,i} \ln(p_{0,i}) - \sum_{i \in S_1} p_{1,i} \ln(p_{1,i}) + \cdots + \sum_{i \in S_n} p_{n,i} \ln(p_{n,i}) = \sum_i L_i(\theta), \tag{2.20}$$

where S_k is the set of indices of the samples from class k, while $p_{k,i}$ is the output probability for class k from sample i [21], [31]. Observing the loss function defined in Equation (2.20) it is clear that in order to calculate its gradient, the gradient of the class probabilities is required, derived using of the chain rule. Using the chain rule once again, these require the neuron values for the current layer, which are found by applying the weights to the previous layer. Hence the need for forward- followed by backpropagation to update the weights of the network.
One challenge when it comes to medical image segmentation is class imbalance in the data. This meansthat the segmentation target, the foreground, is small in comparison to the background. This is known to hamper the training process using conventional cross entropy loss [34]. One manner of addressing this issue is to define multiclass Dice loss function as:

$$L(\theta) = -\frac{2}{|K|} \sum_{k \in K} \frac{\sum_i p_{k,i} g_{k,i}}{\sum_i p_{k,i} + \sum_i g_{k,i}}, \tag{2.21}$$

where k is the class in the set of classes K, $g_{k,i}$ is the ground truth voxel value for class k from sample i and $p_{k,i}$ is as before the output probability of class k for sample i [34].

2.5.2.4 Division of the Training Data

Suitable size of the weight matrices, i.e. the filter size, may differ depending on the problem. If the network were to be trained on all available data, a larger number of weights would always provide a lower loss. This solution will, however, in most cases not be optimal since the network would be more adjusted to the specific examples included in the data set. In some cases, this reduces its ability to generalize the problem, known as overfitting. To avoid this, data is commonly divided into a training set and a validation set. The validation set will be excluded from the weight updating process and will be used only to run through the network in forward passes periodically. If the training loss keeps decreasing while the validation loss instead begins to increase, it is a sign of overfitting [21].
However, when it comes to evaluating the model performance in practice, the validation set is not the most suitable choice. This is preferably done using a separate test set to get as close to the true performance as possible. This data is not to

be involved in any part of the training process and ideally evaluated only once to minimize the risk of overestimation [21].

2.5.2.5 Hyper-Parameters

When training the weights of a neural network, there are certain parameter choices which influence the accuracy of the final model, often referred to as the network's hyper-parameters. These include how many times all examples in the training set are fed to the network, denoted epochs, the chosen learning rate for the optimization and how many examples which are fed to the network simultaneously, called batch size [31].

2.5.2.5.1 Learning Rate

The learning rate, previously called μ, determines the step size of the optimization method. A large value can lead to a faster convergence but sometimes at the cost of never finding the actual minimum getting stuck in an oscillation around it. In contrast, a lower value will be more able to localize the minimum but at the cost of slower convergence. It is therefore common to start at a higher value and successively reduce the learning rate throughout the training process [31].

2.5.2.5.2 Batch Size

In the training process, it is common to not only run one data set through the network at a time, but a random subset of the available examples. How many examples which are used in each set is determined by the batch size parameter [31]. Computing the gradient loss over a large batch of examples enables a closer estimate of the gradient of the entire training. It is also possible that a computation over a batch of size N will be more efficient than N computations over single examples with the use of parallel computational systems [37]. The network accuracy is known to increase the greater the batch size. However, a large batch size comes at a huge computational cost [38].
One commonly implemented way to make the training more efficient is to implement batch normalization [37]. In such cases a normalizing layer is used to normalize the input across the batch and spatial locations. Another approach which in some cases can improve the performance even further is the use of instance normalization instead of batch normalization, where each batch is normalized independently, i.e. solely across spatial locations [39].

2.5.2.5.3 Epoch

An epoch is defined as one random run through of all available training examples. The choice of the total number of training epochs depends also on the choice of learning rate and batch size. It is chosen as an estimate of how many runs of all training examples it should take for the optimization to converge given the assigned learning rate and batch size. Commonly, this number is chosen to be overly large

while introducing an early stop condition concerning the improvement of validation loss. This meaning that if the loss of the validation data set, evaluated at the end of each epoch, has not improved for a certain number of iterations, the training process is aborted [31].

2.5.3 Data Augmentation

The recent progress in performance of CNN:s is largely connected to the development of large data sets and increased computing power. However, the public availability of these large data sets is often limited and training on smaller sets makes the network prone to overfitting [40]. Insufficient amount of available training data is especially common in the medical industry where the data in most cases is heavily protected by privacy regulations. One way to approach this issue is to synthesize new data by warping the original data in different manners. This method is known as data augmentation [41]. This can be done in several different ways and not all of them are suitable for all types of problems. Common examples of augmentation methods for images include: adding noise, change in brightness, reflection and image transformation methods such as rotation and distortion, [21].

3
Methods

In this chapter, the methods developed for this project are presented in three main sections: implementation details regarding the network and training, which data is used and how it is managed and finally the ablation study used to compare different training techniques.

3.1 Implementation Details

In this section the structure of the used network will be presented. Thereafter the implemented augmentation methods will be described followed by more specific details regarding the training procedure of the network. The main network implementation is constructed in Keras, which is a high level neural networks API written in Python, run on top of TensorFlow. The augmentation and data processing is also written in Python and the result visualization is performed in Matlab.

3.1.1 Network Structure

One of the most well known CNN architectures used in medical image analysis is the so called U-net structure first developed in 2D to, later to be extended to 3D. The two main architectural novelties in the U-net were the equal amount of up- and downsampling layers in combination with skip-connections between convolution and deconvolution layers [42]. The 3D version of this network was used as a basis when designing a network for the BRATS 2017 brain segmentation challenge by [34] and this structure is also chosen to be used in this project. The architecture of this network is presented in Figure 3.1.

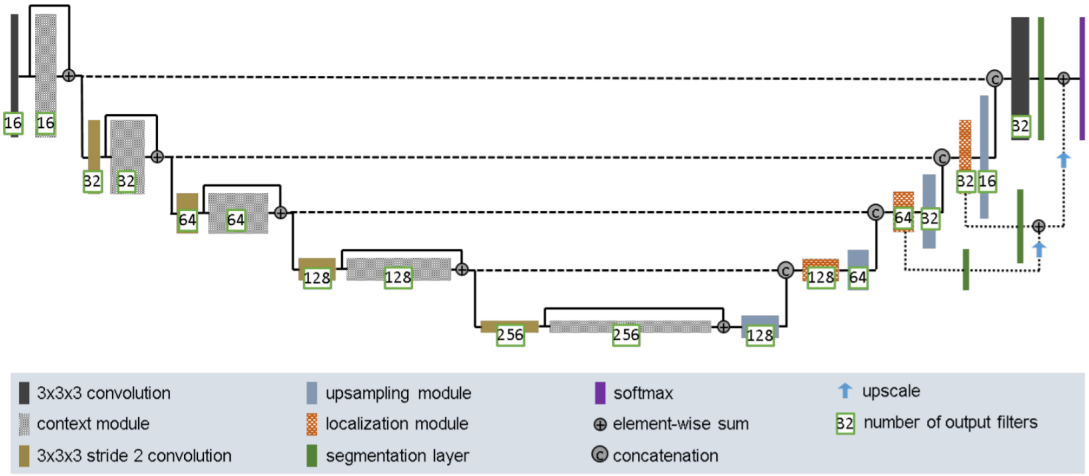


Figure 3.1: This image presents the network architecture used in this project, taken from [34]. It was developed by [34] inspired by the 3D U-net design presented in [43]. The left side of this structure is referred to as the context pathway and the right the localization pathway [34].

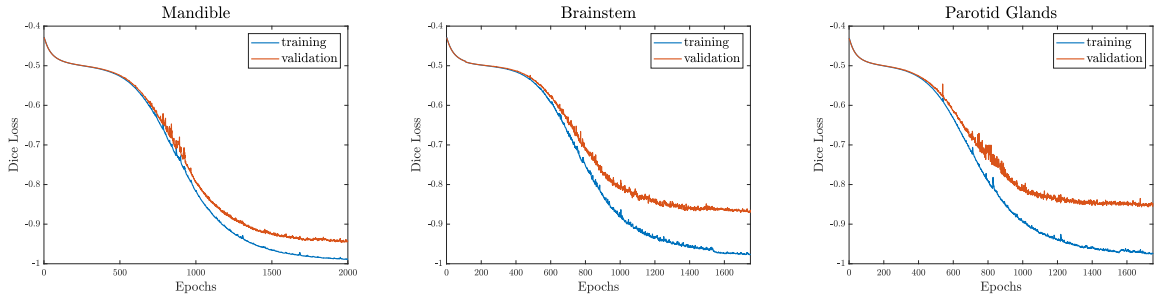
Similarly to U-net, this design uses a context aggregation pathway to encode increasingly abstract input representations progressing through the network followed by a localization pathway recombining these with shallower input. This, giving the network the characteristic U-shaped appearance. Difference in design compared to 3D U-net include the specific architecture of the context pathway, the normalization process, the number of network feature maps, non-linearity and the structure of the localization pathway [34].
In this network, the context modules are described as pre-activation residual blocks consisting of two $3 \times 3 \times 3$ convolutional layers and a dropout layer with rate 0.3 in between [34]. This differs from the original U-net where simple convolution layers followed by max pooling are used to reduce the feature map dimension [43].
In the localization pathway, feature maps are initially upsampled and upscaled, followed by a $3 \times 3 \times 3$ convolution. These are thereafter recombined with features at the same level through concatenation whereafter a localization module is used to combine the features. The implemented localization module consist of a $3 \times 3 \times 3$ convolution followed by a $1 \times 1 \times 1$ convolution [34]. By significantly reducing the number of feature maps in this pathway, this network is able to train twice as many filters as the original U-net with only a slightly smaller image size and also a larger batch size [34].
Deep supervision is employed by adding segmentation layers in the localization pathway at different levels which later are added to the final network output. In all feature map convolutions throughout the network, a leaky ReLU nonlinearities are used with a negative slope of 10^{-2}. While 3D U-net suggests batch normalization, this implementation uses instead instance normalization. This, since the stochasticity induced by using a relatively small batch size might destabilize this method [34].

3.1.2 Augmentation Methods

Dealing with medical imaging, such as CT, both the technique and the image motif are in this case taken into consideration when determining relevant means of augmentation. In this project three augmentation methods are implemented: distortion, rotation and reflection. These are deemed to be the most relevant for this particular task.
Small distortions can be used to represent the different anatomies between patients. Reflection along the sagittal plane can be relevant to reflect natural individual asymmetries but reflection around any other plane is considered unrealistic. The patient position within the CT-machine is not entirely fixed which means that rotational augmentation is relevant. However, the mobility range is limited which in turn limits the range of angles relevant for this augmentation.
Each data set is paired with its own 4×4 affine transformation matrix, a combination of the rotation and translation portion, which maps the voxel coordinates to the spatial coordinate system. The distortion is derived applying randomly selected scale factors for each dimension in the image plane to its affine matrix and transforming the image and its label map accordingly. The scale factors are derived from a normal distribution with mean 1 and a chosen standard deviation. Similarly, the rotation is applied by rotating the image affine using of a rotation matrix defined as in Equation (2.3), where θ_x, θ_y, and θ_z are random from a normal distribution with mean 0 and a chosen standard deviation. The reflection is performed by simply reverting the order of the image planes in the x-dimension, which switches the order of the patient's lateral left and right. This is performed according to a uniformly distributed random boolean. When augmentation was employed, it was executed on all examples in between each epoch. This, resulting in the network being exposed to slightly different data sets each epoch thus simulating the existence of a larger amount of available examples.

3.1.3 Training Procedure

The network was trained alternating between two NVIDIA Titan X GPU:s, one GetForce GTX and one Pascal, where the number of epochs where determined with a patience of 200 epochs of unimproved validation loss. The graphic cards were used separately since there was an unresolved multithreading issue connected to the data handling system. Training was, as in [34], executed using the Adam optimizer with an initial learning rate of $5 \cdot 10^{-4}$ for a weighted dice loss as defined in Equation (2.21). If there was no improvement of validation loss for the last 100 epochs, the learning rate was reduced by a factor 0.5. When data augmentation was employed, it was executed for each training example but the validation data was kept untouched. The batch size was chosen to be as large as possible and it was limited by the GPU memory to a maximum of 2.
When choosing standard deviation for the distortional and rotational augmentation methods, they were initially applied separately to a data set to empirically determine visually feasible factors. Observing the training process of the networks with no augmentation there was an apparent variation in the difference of training and validation set loss convergence for the different organs, as illustrated in Figure 3.2



(a) **(b)** **(c)**

Figure 3.2: Subfigures 3.2a-3.2c present the loss convergence for the training set compared to the validation set for the mandible, brainstem and the parotid glands, respectively.

The goal was to reduce the gap between the training and validation loss as much as possible by the use of augmentation to minimize overfitting. Meaning that a larger gap between the curves would require larger augmentation parameters due to a larger difference between the data sets. This process resulted in the use of the standard deviations presented in Table 3.1.

Table 3.1: This table presents the different standard deviations $\sigma_{distortion}$ and $\sigma_{rotation}$ used for data augmentation when training a convolutional neural network to segment three different organs: the mandible, the brainstem and the parotid glands.

	$\sigma_{distortion}$	$\sigma_{rotation}$
Mandible	0.05	0.005
Brainstem	0.10	0.010
Parotid Glands	0.15	0.015

To achieve as high level of comparability as possible to the results obtained in the MICCAI 2015 challenge, the division of the data into a training, validation and test set was chosen according to their standards. Where 25 data sets were used training data, the 10 off-site test sets was used for validation and the 5 on-site sets were completely excluded from the training process and used solely for final testing.

3.2 Data Specification

Every year different so called grand challenges within the field of biomedical image analysis are organized. In each, a certain medical image analysis task is put fourth along with image data and a set of conditions. This enables full comparability of different state-of-art approaches.
The goal of the MICCAI 2015 Head and Neck Auto Segmentation Challenge was to compare the performance of different automatic segmentation techniques to be used for treatment planning of radiation oncology. Delineation of target structures, such as the tumor and organs at risk, is a key step of this process. As in hyperthermia

treatment planning, the fact that manual segmentation is highly time consuming is a limiting factor, therefore the need for an effective automatic segmentation method [1].
There were in total 6 participating teams called FH, IM, UB, UC, UW and VU. Of the submitted segmentation approaches to the MICCAI 2015 challenge, 3 were atlas based (FH, UC and VU), one used a model based approach by active appearance modelling (IM), one used combined atlas and model based with active shape modelling (UB) and one used basic image processing by landmark detection (UW). The IM team was announced winner, UB got the second place and VU third.

3.2.1 Data Description

Originally, the data used in this challenge comes from a clinical trial with CT scans of 111 patients for treatment planning. This data is publicly available and from this a subset of 40 images was chosen for the challenge for quality reasons, 25 of which was to be used for training, 10 for off-site testing and 5 for on-site testing [1]. There were 8 optional additional training cases also provided, which were not fully annotated. In addition, there were cases where the training data lacked annotation of one of the organ types.
These images were manually segmented by experts to provide a uniform segmentation quality and consistency. Guidelines for this were developed during an extensive literature research and this procedure was performed for 9 structures considered to be organs at risk: the brainstem, optic chiasm, mandible, left and right optic nerves, parotid glands and submandibular glands, illustrated in Figures 1.1 and 1.2 [1].
All provided images are of size 512×512 in the xy-plane and the in-plane pixel resolution is isotropic and varied in between 0.76 mm$\times$0.76 mm to 1.27 mm $\times$ 1.27 mm [1]. The number of slices varied in the range of 76-360 and the slice spacing varies in between 1.25 mm and 3 mm.

3.2.2 Data Management

The used network is designed to process 3D blocks of dimension $128 \times 128 \times 128$ [34]. Two different methods of preprocessing the input data to this shape were implemented and the resulting network performance was compared. In one the full scale image was cropped according to the organ, foreground, location and resized by interpolation. All images are of the same size in the xy-dimension, 512×512 pixels, but the z-dimension varied. In order for all image sets to have the same voxel spacing, they were padded to the same size of 360 before this resizing process. The second method consisted in training the network using patches of the full resolution image of size $128 \times 128 \times 128$. This consumed more process memory since the data sets uploaded were significantly larger, which required downsampling by a factor 2. In both cases, all image sets were normalized to values between 0 and 1.
As a final step to the segmentation, the output prediction of the organ segmentation was post-processed by removing connected component of a volume below a certain threshold. This was done to avoid outlier points from affecting the final result.

3.3 Ablation Study

To investigate the impact of different training choices, apart from hyper-parameter values, an ablation study is performed. The following scenarios are investigated separately for one organ, the mandible:

- Training of the whole image resized to $128 \times 128 \times 128$.
- Training on the image downsampled by 2 in with patches of size $128 \times 128 \times 128$.
- With reflection augmentation.
- With distortion augmentation.
- With rotation augmentation.
- All augmentations combined.

Initially, training on the full image resized to $128 \times 128 \times 128$ is compared to training on the downsampled image with patches of size $128 \times 128 \times 128$. The choice proven to be the most advantageous in respect to the quality indicators presented in Section 2.4.2 is further used in the testing of the different augmentation methods. After training the network with the different augmentation methods, the performance of these are evaluated by the use of the same metrics. The augmentation method which performs the highest is chosen to be used to train the network for the remaining organs.
For each training process, a model is chosen according to the highest over all dice value for the validation set. This model is thereafter used to predict a segmentation of the test set, in respect to all metric values. The performance in the test set is the main basis for evaluating the superiority of each method tested.

4

Results

When training the network the loss generally decayed according to the shapes of the curves presented in Figure 4.1, although convergence was achieved at different levels and rates depending on the amount of data, augmentation and organ type. Figure 4.1 presents an example of the difference in loss convergence when training on the whole image, patches of the image and different standard deviations σ for the same augmentation type. Note that the validation loss is lower for the augmented cases, Subfigure 4.1c and 4.1d, compared to the case with no augmentation, Subfigure 4.1a. This was generally found to be the case for all organs. A model trained using the augmentation factor which generated a loss convergence as in Subfigure 4.1c was found to not perform well as the model which was generated with the loss convergence as in 4.1d.

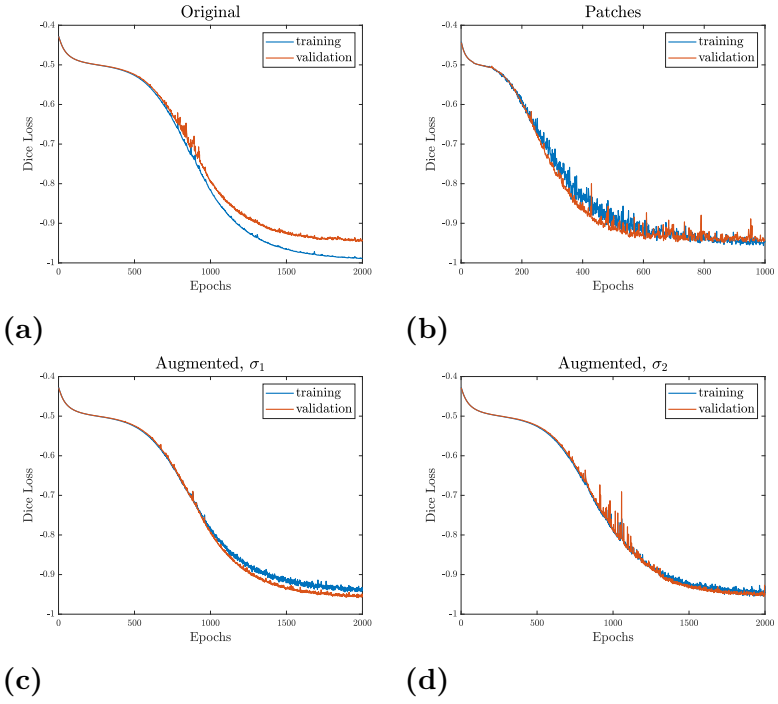


Figure 4.1: This figure illustrates the difference in loss decay for four different training scenarios: *original*, *patches* and *augmented*, where the image was augmented using the same method but with different standard deviations for the augmentation parameter σ_1 and σ_2, $\sigma_1 > \sigma_2$. Here, *original* represents training on the whole image and *patches* training on patches of the image. All plots were obtained during segmentation training for the mandible.

There were several factors influencing the time it took to train the network, where the main factors appeared to be the type of graphics card and the amount of available data. For the mandible, with a total of 25 available training sets and 10 validation sets, training on the whole image resized to $128 \times 128 \times 128$ led to an epoch training time of 25 s on the Pascal card and 38 s on the GetForce GTX. When training on patches of the downsampled image the amount of data was consequently larger and the training took 104 s on the Pascal card and 124 s on the GetForce GTX per epoch. Both the brainstem and the parotid glands had 33 available training data sets and 10 validation data sets. These organs were only trained on whole resized images, and one epoch took 32 s on the Pascal and 47 s on the GetForce GTX.
Convergence occurred after around 2000 epochs for training segmentation of the mandible on the whole resized image and after around 1000 epochs when training with patches of the downsampled image. For the brainstem and the parotid glands, convergence occurred after around 1750 epochs. Thus resulting in the total training time as presented under *Training* in Table 4.1. This table also present the time it took to one data set using the trained model for segmentation of the whole image at once and using patchwise segmentation. Note that the training time is presented in hours and the segmentation time in seconds.

Table 4.1: This table presents the total amount of time it took to train the model compared to segmentation of one data set using the trained model for different organs and training modes. The two values in each column represent the use of two different editions of Titan X graphics cards: Pascal to the left and GetForce GTX to the right.

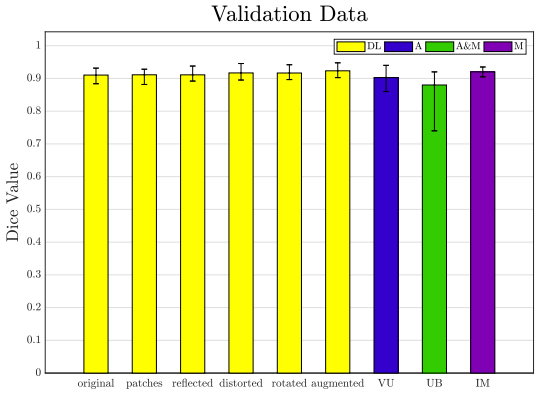
	Training [h]	Segmenting [s]
Mandible - Whole Image	13.88/21.11	1.9/2.1
Mandible - Patches	28.89/34.44	19/21
Brainstem Parotid Glands	15.56/22.85	1.9/2.1

The following sections presented the results obtained when training the network for the three different organs: the mandible, brainstem and the parotid glands. This is followed by a presentation of a final combined segmentation containing all three organs compared to the ground truth.

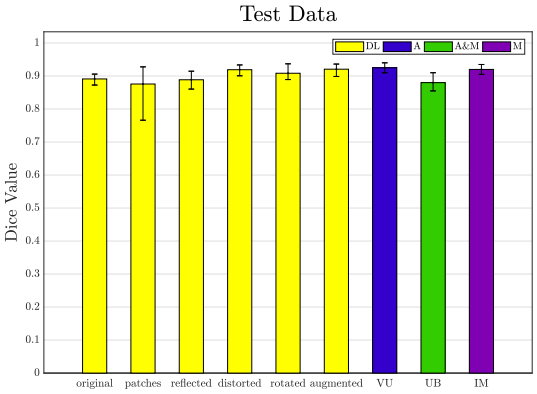
4.1 The Mandible

After training the network to perform segmentation of the mandible for all 6 scenarios presented in Section 3.3, the performance of the resulting models were evaluated on both the validation and test set in terms of the Dice coefficient values, the 95th percentile Hausdorff distance (HD), the maximum HD and the contour mean distance (CM), see definition in Section 2.4.2. The resulting values are presented in the bar charts in Figure 4.2. In this same figure, the obtained values are also put

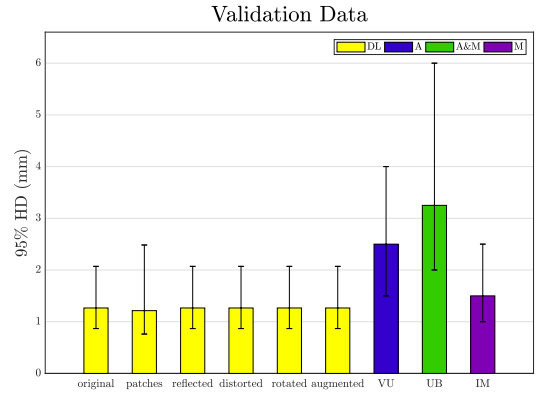
in relation to the values of the top three teams in the MICCAI 2015 challenge, IM, UB and VU, approximated from the bar charts presented in [1].



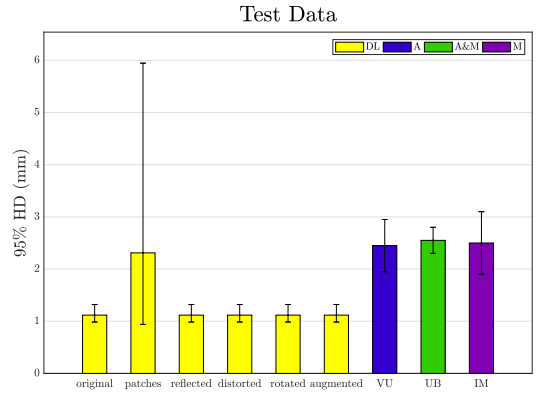
(a)



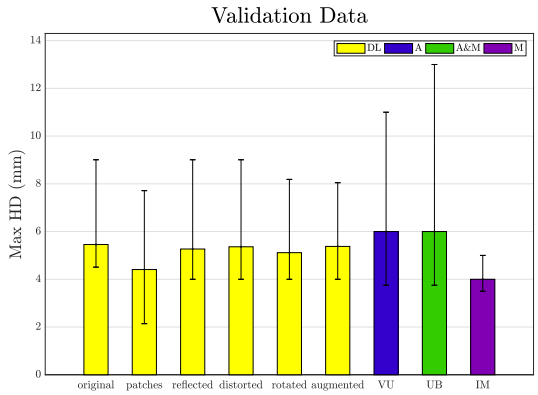
(b)



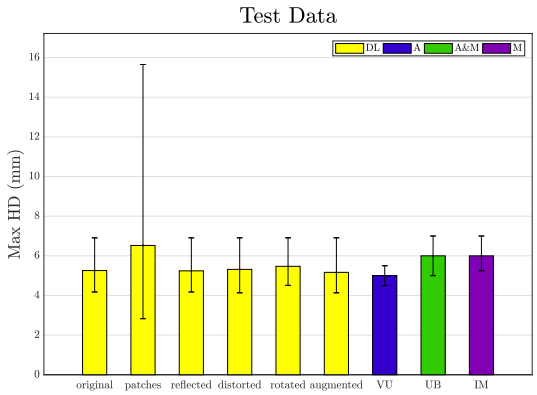
(c)



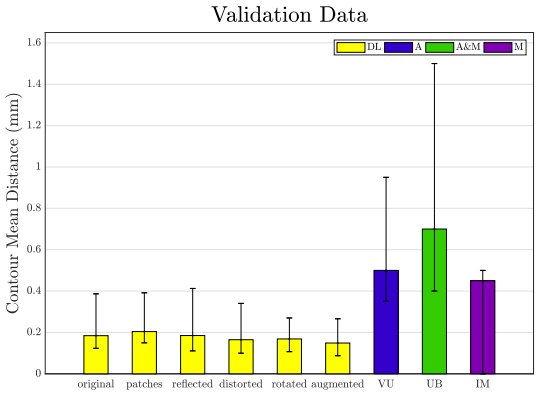
(d)



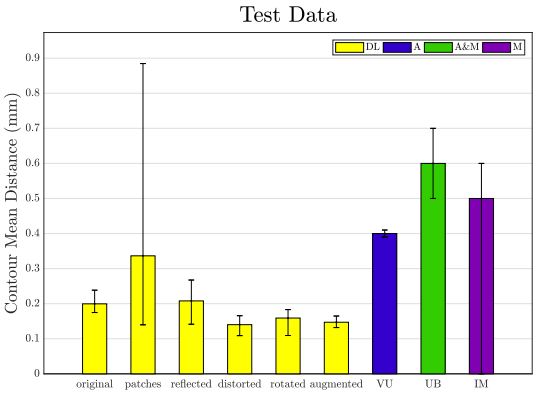
(e)



(f)



(g)



(h)

Figure 4.2: The bar charts in Subfigures 4.2a-4.2h present the metric values obtained during the ablation study for the mandible compared to the values of the top three teams, *IM*, *UB* and *VU*, participating in the MICCAI 2015 challenge. Their values have been estimated from charts presented in [1]. The bar height represents the mean parameter value of the evaluated data set and the whiskers represent the 5th and 95th percentile respectively. On the x-axis, *original* and *patches* represent training on the whole resized image and training on patches of higher resolution images of the same size as the resized image, respectively. *reflected*, *distorted* and *rotated* represent the cases where only these augmentation types where used, respectively, and *augmented* represent the case where all three methods were employed simultaneously. *DL* stands for deep learning, *A* for atlas based, *A&M* for atlas and model based and *M* for model based segmentation.

When observing the bar charts in Figure 4.2, note that when training on patches of the image affected the performance distinctively in a disadvantageous sense when evaluating the segmentation metrics on the test set. Disadvantageous, given that the dice value is on average lower and the remaining surface parameters on average higher with a larger uncertainty across the test set. Figure 4.3 displays a 3D representation of the data set which generated the highest and lowest dice values across the test set for the network trained on patches of the downsampled image.

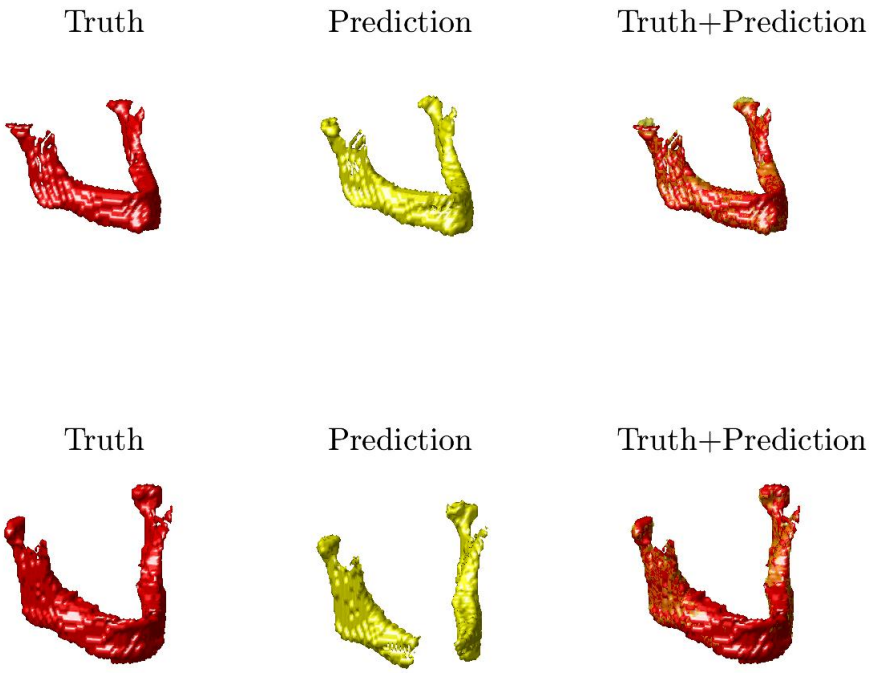
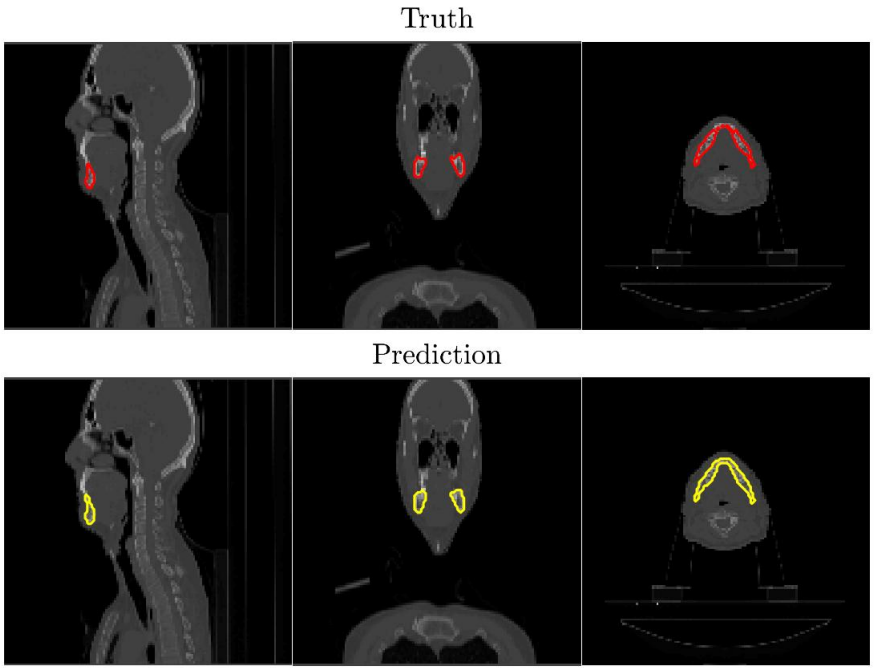
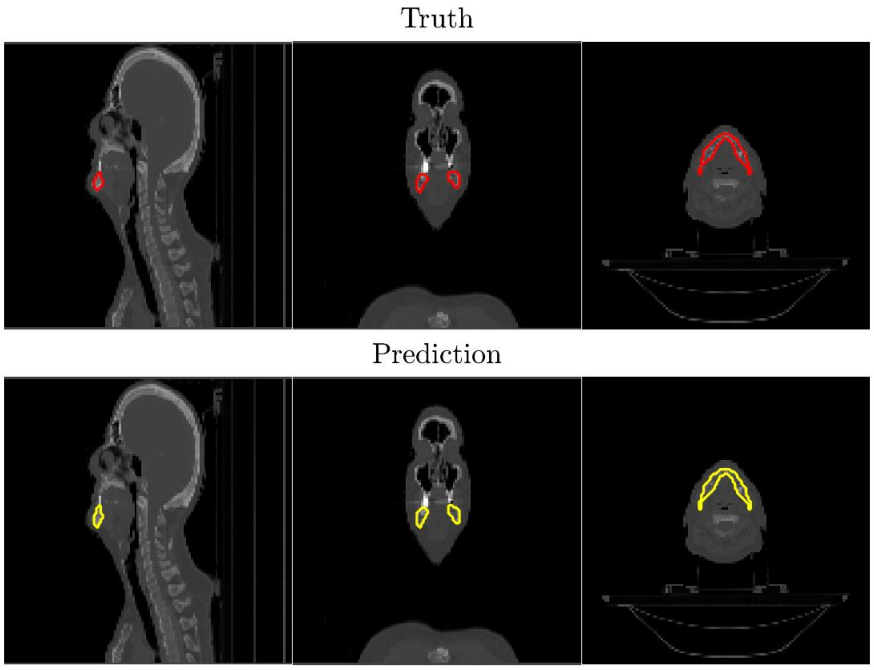


Figure 4.3: Subfigures 4.3a and 4.3b display 3D models of the mandible comparing the ground truth to the predicted segmentation with the highest and lowest acquired Dice value for the test set of 0.766 and 0.928, respectively, when training the network with patches of the downsampled image. The *Truth+Prediction* is constructed by overlaying a tranparent version of the prediction on top of the ground truth.

Figure 4.4 presents a 2D representation of the prediction compared to the ground truth for the test data set which obtained the highest and the lowest dice value for the entire test set for the case found to perform the highest in respect to the metric values presented in Figure 4.2: when all augmentation methods were employed simultaneously.



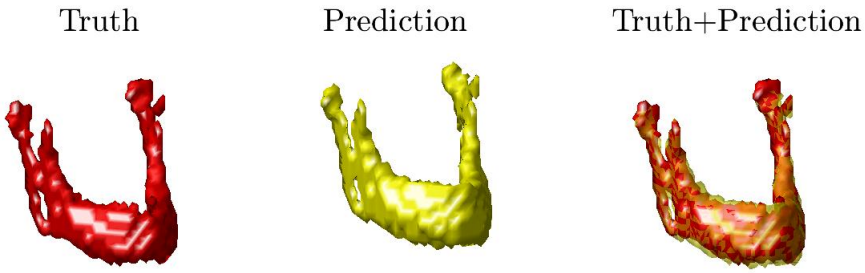
(a)



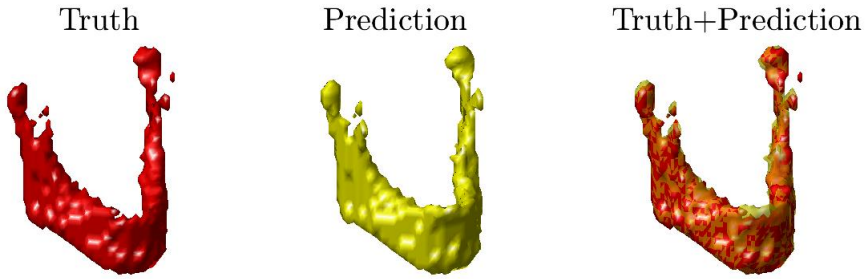
(b)

Figure 4.4: Subfigure 4.4a presents the median plane of the mandible in all 3 dimensions comparing the ground truth to the segmentation with the highest acquired Dice value for the test set of 0.936 and Subfigure 4.4b comparing the ground truth to the predicted segmentation with the lowest Dice value of 0.899. The network was trained using reflection, distortion and rotation augmentation.

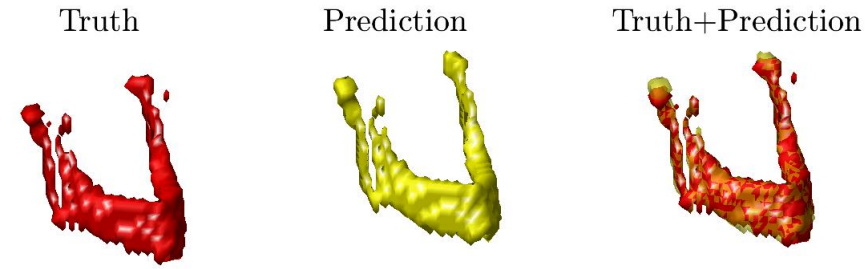
Finally, Figure 4.5 illustrates the segmentation of the entire test set with the corresponding Dice values of 0.8987, 0.936, 0.926, 0.929 and 0.913. Note that the ground truth is not fully connected in Subfigures 4.5b-4.5d. This is an artifact from the resizing of the image in the preprocessing stage.



(a)



(b)



(c)

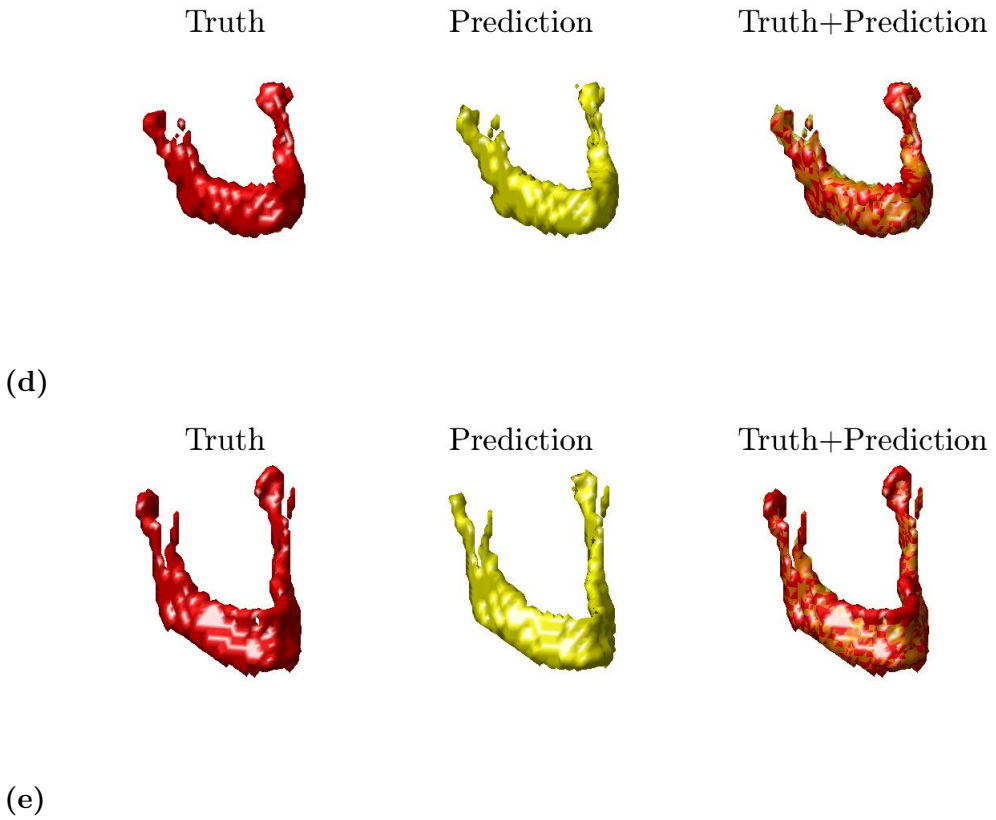
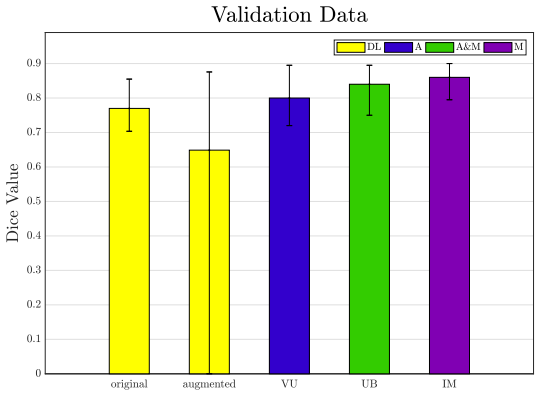


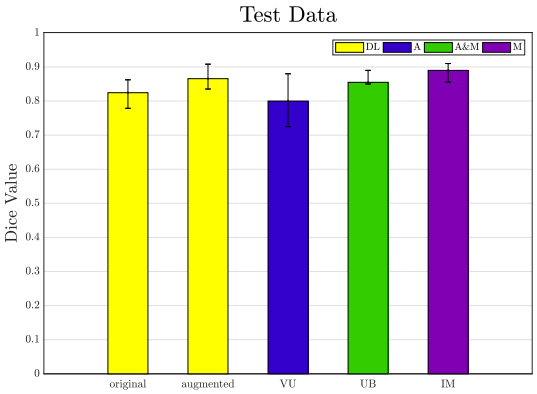
Figure 4.5: Subfigures 4.5a-4.5e display 3D models of the mandible comparing the ground truth to the predicted segmentation for the entire test set, with corresponding Dice values of: 0.899, 0.936, 0.926, 0.929 and 0.913. The *Truth+Prediction* is constructed by overlaying a transparent version of the prediction on top of the ground truth. The network was trained using reflection, distortion and rotation augmentation.

4.2 The Brainstem

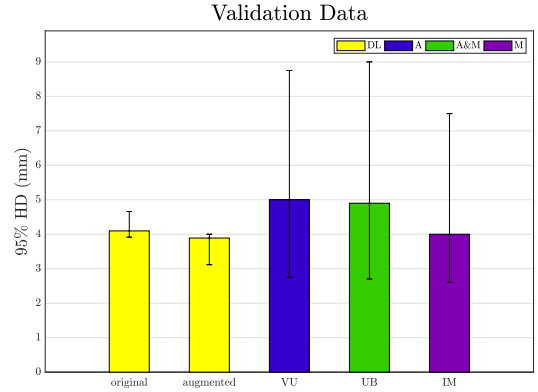
As presented in Section 4.1, the augmentation method found to be the most advantageous in respect to the metrics presented in Section 2.4.2 was the case where reflection, distortion and rotation was applied simultaneously. This was therefore the augmentation method used to train the network for segmentation of the remaining two organs as well. Figure 4.6 present bar charts of the evaluation metrics for the brainstem segmentation for the case with no augmentation and using all three augmentation methods. Also in this case, the values obtained are put in relation to the top three teams in the MICCAI 2015 challenge. For the augmented case, the final model missed labeling the organ in 2 out of 10 data sets, affecting the Dice coefficient for this case in Subfigure 4.6a prominently. For the Hausdorff distance and contour mean distance, they were excluded from the statistics presented in Subfigures 4.6c, 4.6e and 4.6g since it is not possible to calculate a surface distance without a surface. Note that this should be kept in mind when comparing the performance of the different segmentation options in respect to this data set.


Validation Data
DL
A
A&M
M
Dice Value
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
original
augmented
VU
UB
IM

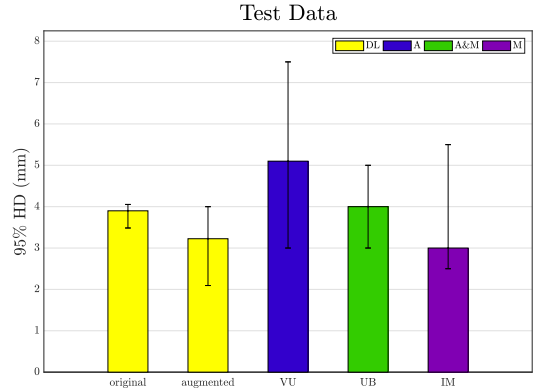
(a)


Test Data
DL
A
A&M
M
Dice Value
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
original
augmented
VU
UB
IM

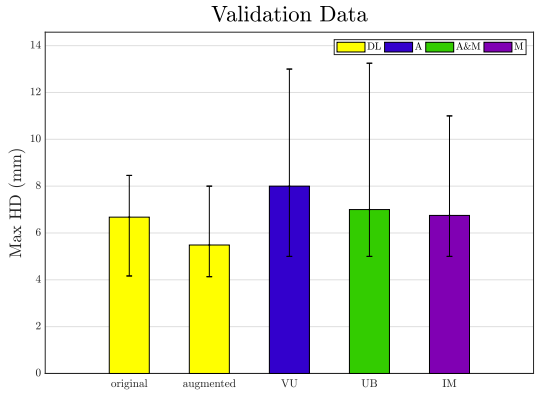
(b)


Validation Data
DL
A
A&M
M
95% HD (mm)
0
1
2
3
4
5
6
7
8
9
original
augmented
VU
UB
IM

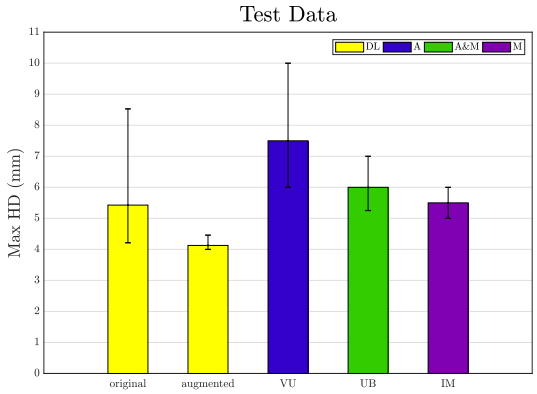
(c)


Test Data
DL
A
A&M
M
95% HD (mm)
0
1
2
3
4
5
6
7
8
original
augmented
VU
UB
IM

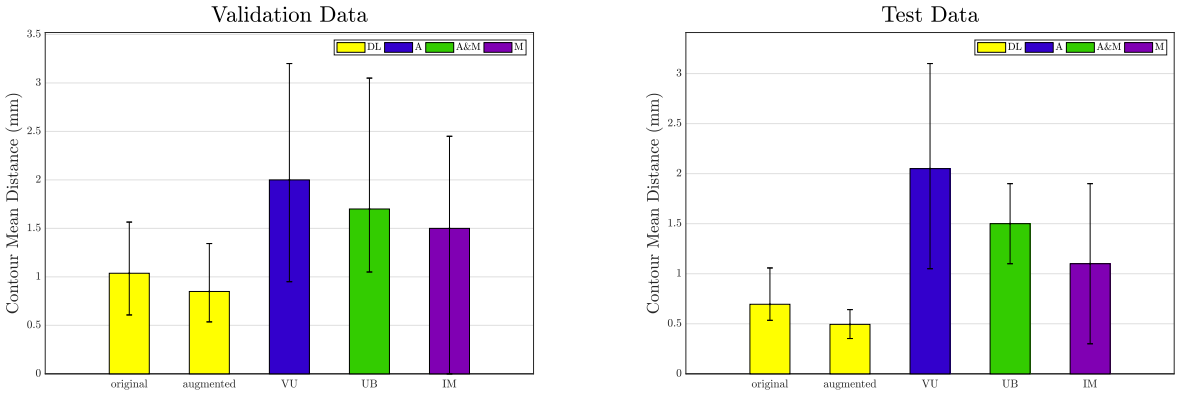
(d)


Validation Data
DL
A
A&M
M
Max HD (mm)
0
2
4
6
8
10
12
14
original
augmented
VU
UB
IM

(e)


Test Data
DL
A
A&M
M
Max HD (mm)
0
1
2
3
4
5
6
7
8
9
10
11
original
augmented
VU
UB
IM

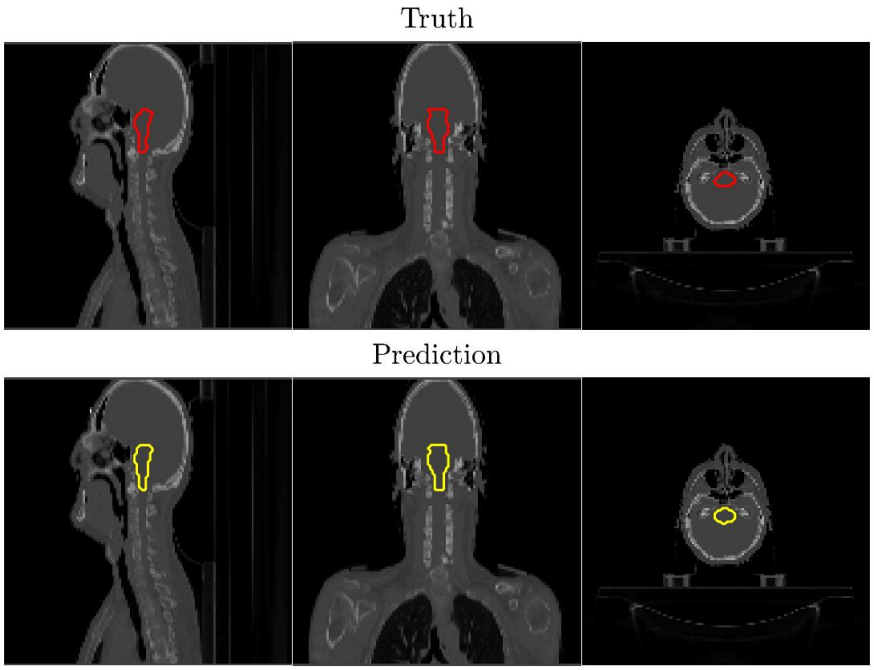
(f)



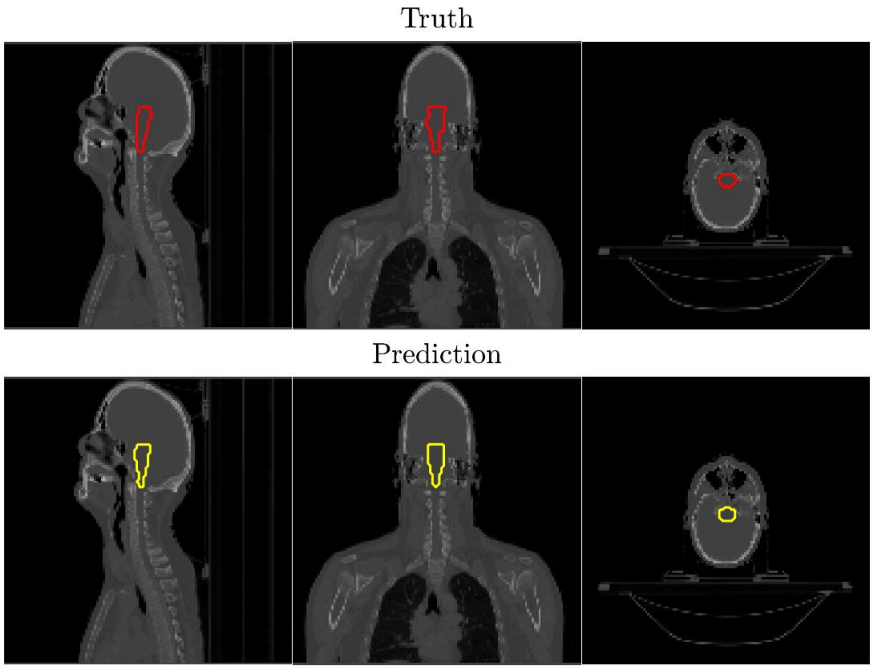
(g) **(h)**

Figure 4.6: The bar charts in Subfigures 4.6a-4.6h present the metric values obtained for the brainstem compared to the values of the top three teams, *IM*, *UB* and *VU*, participating in the MICCAI 2015 challenge. Their values have been estimated from charts presented in [1]. The bar height represents the mean parameter value of the evaluated data set and the whiskers represent the 5th and 95th percentile respectively. On the x-axis. *original* represent training on the whole resized image without augmentation and *augmented* represent the case where all three methods were employed simultaneously. *DL* stands for deep learning, *A* for atlas based, *A&M* for atlas and model based and *M* for model based segmentation. In the augmented case for the validation set, the organ was missed 2 out of 10 times affecting the dice parameter prominently. These examples were excluded from the statistics in Subfigures 4.6c, 4.6e and 4.6g.

Once again, the augmented case generated the highest performance on the test set in respect to all evaluation metrics observed in Figure 4.6. Figures 4.7a present 2D representation of the predictions for the data set with both the highest and lowest obtained dice value. Note that in Subfigure 4.7b, it is apparent that the model has failed to capture the asymmetry found in the corresponding ground truth in respect to the anatomical symmetry axis. Finally, Figure 4.8 present the predicted segmentation of the brainstem in the entire test set with the corresponding Dice values of 0.873, 0.852, 0.860, 0.835 and 0.908.



(a)

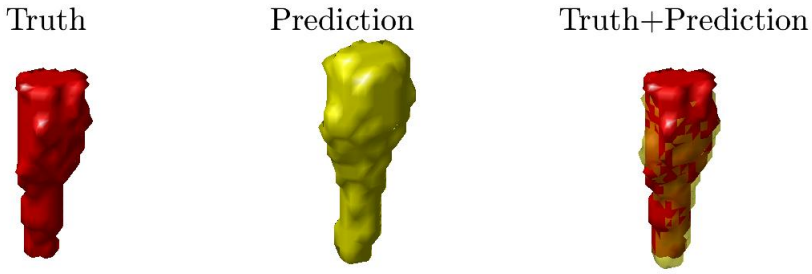


(b)

Figure 4.7: Subfigure 4.7a present the median plane of the brainstem in all 3 dimensions comparing the ground truth to the segmentation with the highest acquired Dice value for the test set of 0.908 and Subfigure 4.7b comparing the ground truth to the predicted segmentation with the lowest Dice value of 0.835. The network was trained using reflection, distortion and rotation augmentation.


Truth
Prediction
Truth+Prediction

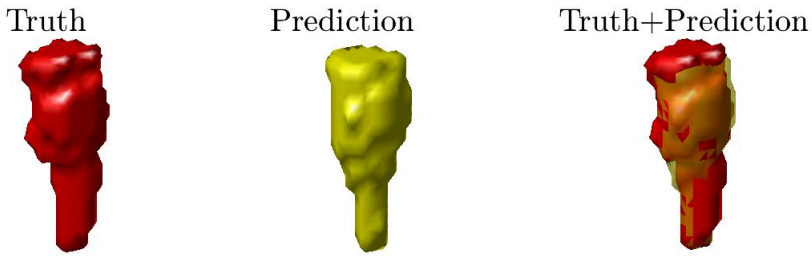
(a)


Truth
Prediction
Truth+Prediction

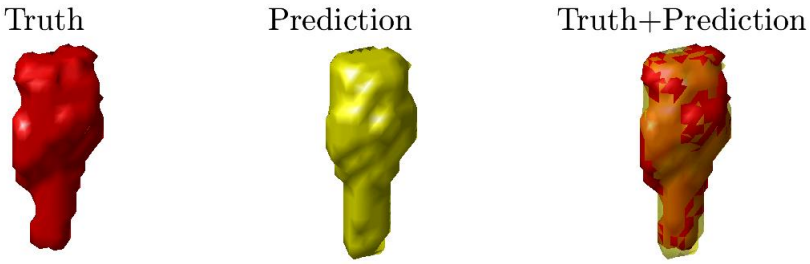
(b)


Truth
Prediction
Truth+Prediction

(c)


Truth
Prediction
Truth+Prediction

(d)

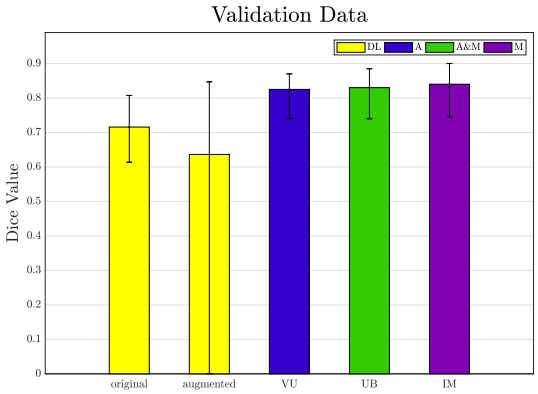


(e)

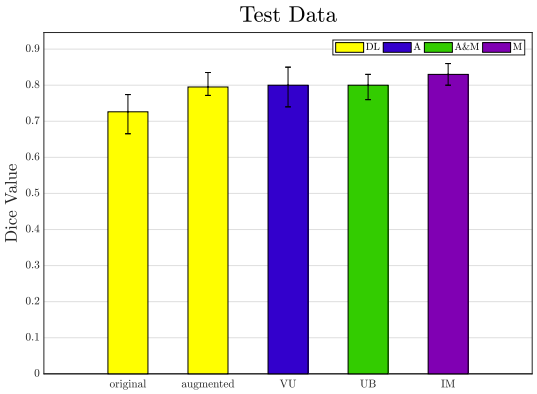
Figure 4.8: Subfigures 4.8a-4.8e displays 3D models of the brainstem comparing the ground truth to the predicted segmentation for the entire test set, with corresponding Dice values of: 0.873, 0.852, 0.860, 0.835 and 0.908. The *Truth+Prediction* is constructed by overlaying a transparent version of the prediction on top of the ground truth. The network was trained using reflection, distortion and rotation augmentation.

4.3 The Parotid Glands

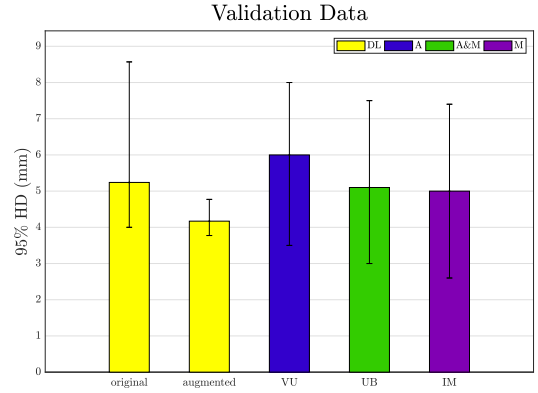
Also here, the augmentation method used was a combination of reflection, distortion and rotation. Figure 4.9 present bar charts of the evaluation metrics, presented in Section 2.4.2, obtained when segmenting the parotid glands in the validation and test set. Here, the results obtained using augmentation is compared to the case with no augmentation, marked as *original*, and the results obtained by the top three teams of the MICCAI 2015 challenge: IM, UB and VU. For the augmented case, the final model missed labeling the organ in 2 out of 10 data sets, affecting the Dice coefficient for this case in Subfigure 4.9a prominently. For the Hausdorff distance and contour mean distance, they were excluded from the statistics presented in Subfigures 4.9c, 4.9e and 4.9g since it is not possible to calculate a surface distance without a surface. Note that this should be kept in mind when comparing the performance of the different segmentation options in respect to this data set.


Validation Data
DL
A
A&M
M
Dice Value
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
original
augmented
VU
UB
IM

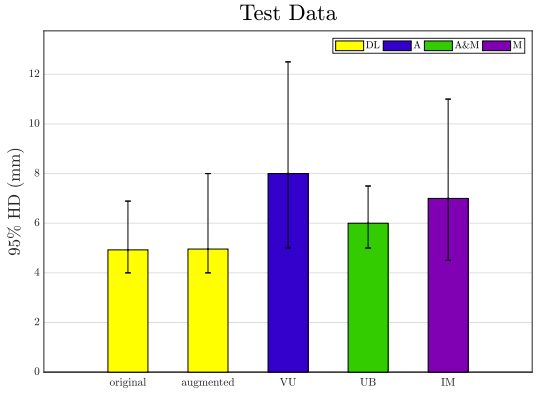
(a)


Test Data
DL
A
A&M
M
Dice Value
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
original
augmented
VU
UB
IM

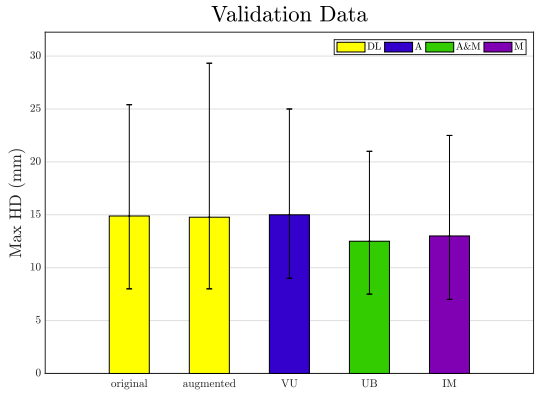
(b)


Validation Data
DL
A
A&M
M
95% HD (mm)
0
1
2
3
4
5
6
7
8
9
original
augmented
VU
UB
IM

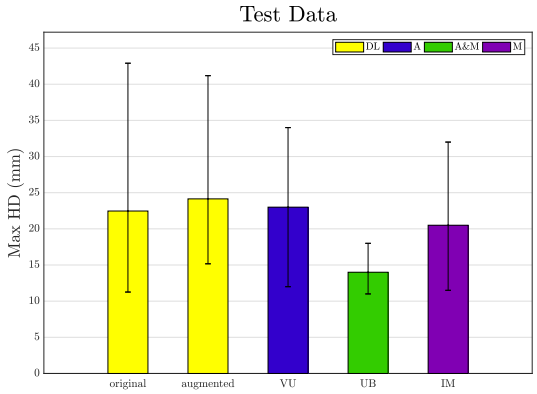
(c)


Test Data
DL
A
A&M
M
95% HD (mm)
0
2
4
6
8
10
12
original
augmented
VU
UB
IM

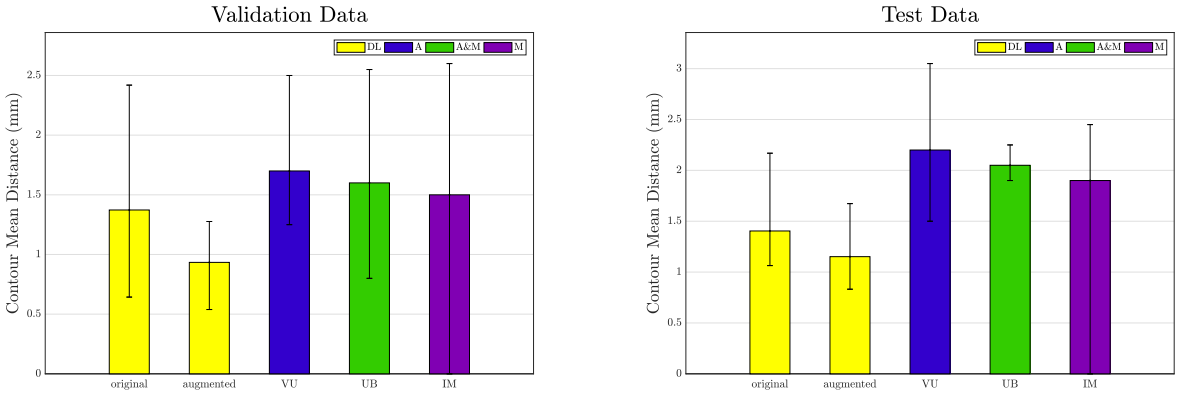
(d)


Validation Data
DL
A
A&M
M
Max HD (mm)
0
5
10
15
20
25
30
original
augmented
VU
UB
IM

(e)


Test Data
DL
A
A&M
M
Max HD (mm)
0
5
10
15
20
25
30
35
40
45
original
augmented
VU
UB
IM

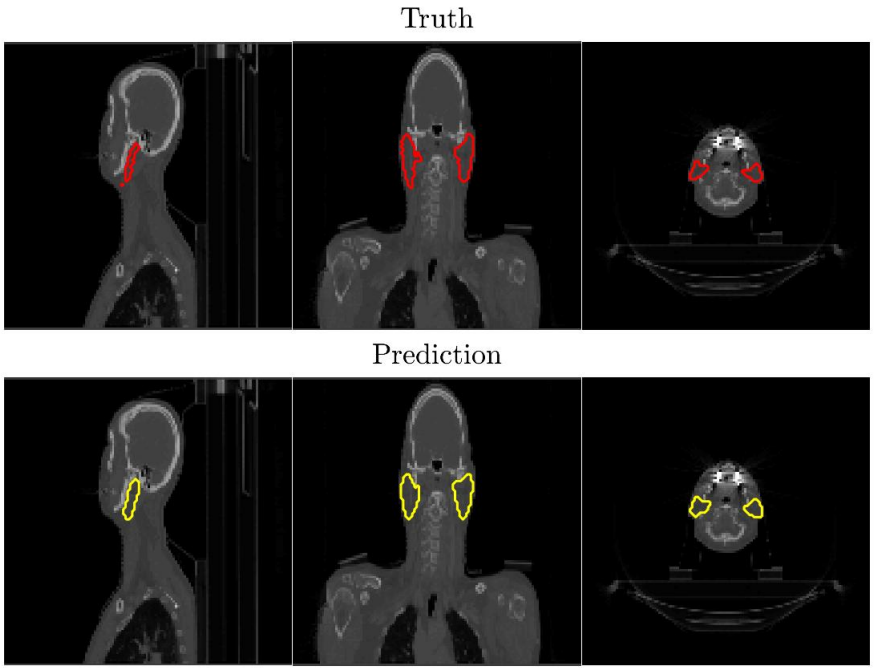
(f)



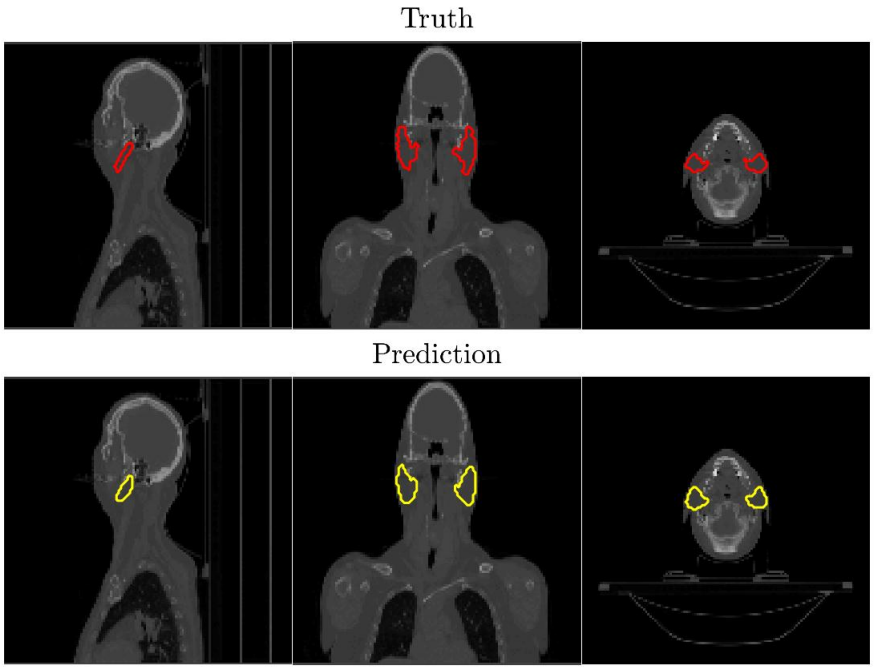
(g) **(h)**

Figure 4.9: The bar charts in Subfigures 4.9a-4.9h present the metric values obtained for the parotid glands compared to the values of the top three teams, *IM*, *UB* and *VU*, participating in the MICCAI 2015 challenge. Their values have been estimated from charts presented in [1]. The bar height represents the mean parameter value of the evaluated data set and the whiskers represent the 5th and 95th percentile respectively. On the x-axis. *original* represent training on the whole resized image without augmentation and *augmented* represent the case where all three methods were employed simultaneously. *DL* stands for deep learning, *A* for atlas based, *A&M* for atlas and model based and *M* for model based segmentation. In the augmented case for the validation set, the organ was missed 2 out of 10 times affecting the dice parameter prominently. These examples were excluded from the statistics in Subfigures 4.9c, 4.9e and 4.9g.

Figure 4.10 presents a 2D representation in three orthogonal planes of the test data sets in which the model obtained the highest and lowest Dice values. These results were obtained using the model which had the highest performance in respect to the metrics presented in Figure 4.9 for the test set: the network which was trained using data augmentation. A 3D representation of the prediced segmentation of the parotid glands for the entire test set is presented in Figure 4.11. As for the mandible, note that in Subfigure 4.11b the ground truth is not fully connected due to an artifact occurring during the resizing process of the original data. Overall when observing Subfigured 4.11a-4.11e, it is worth noting that there is a wide variation in the shape of the glands in the ground truth which is not fully captured in the predicted segmentation.

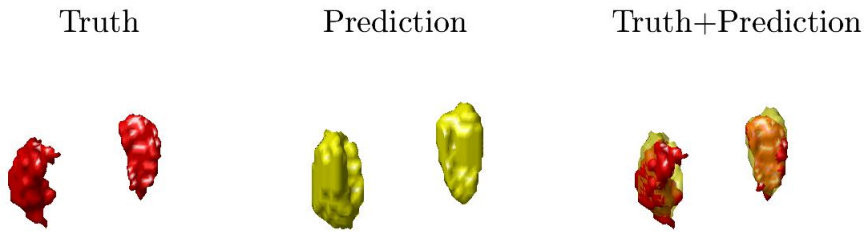


(a)

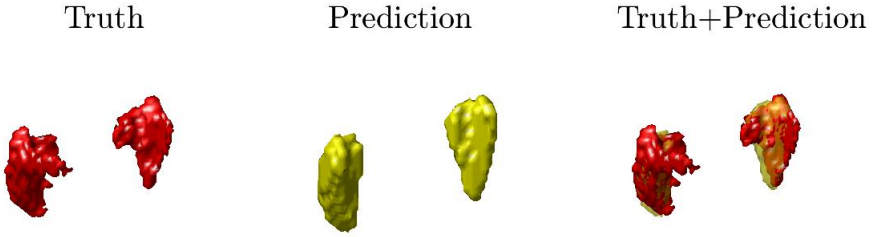


(b)

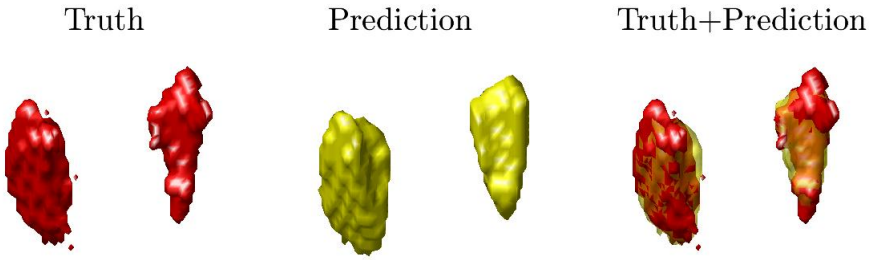
Figure 4.10: Subfigure 4.10a presents the median plane of the parotid glands in all 3 dimensions comparing the ground truth to the segmentation with the highest acquired Dice value for the test set of 0.835 and Subfigure 4.10b comparing the ground truth to the predicted segmentation with the lowest Dice value of 0.771. The network was trained using reflection, distortion and rotation augmentation.


Truth
Prediction
Truth+Prediction

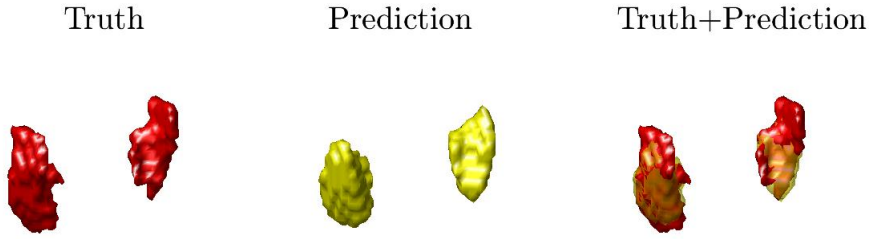
(a)


Truth
Prediction
Truth+Prediction

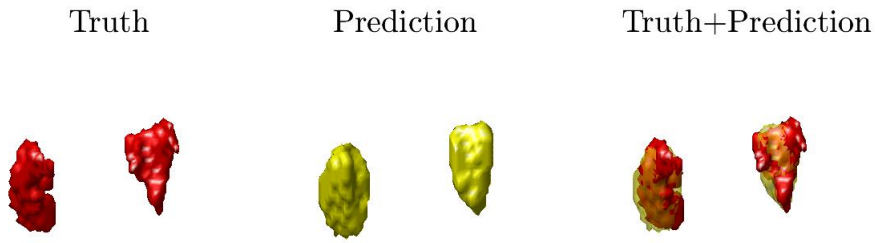
(b)


Truth
Prediction
Truth+Prediction

(c)


Truth
Prediction
Truth+Prediction

(d)

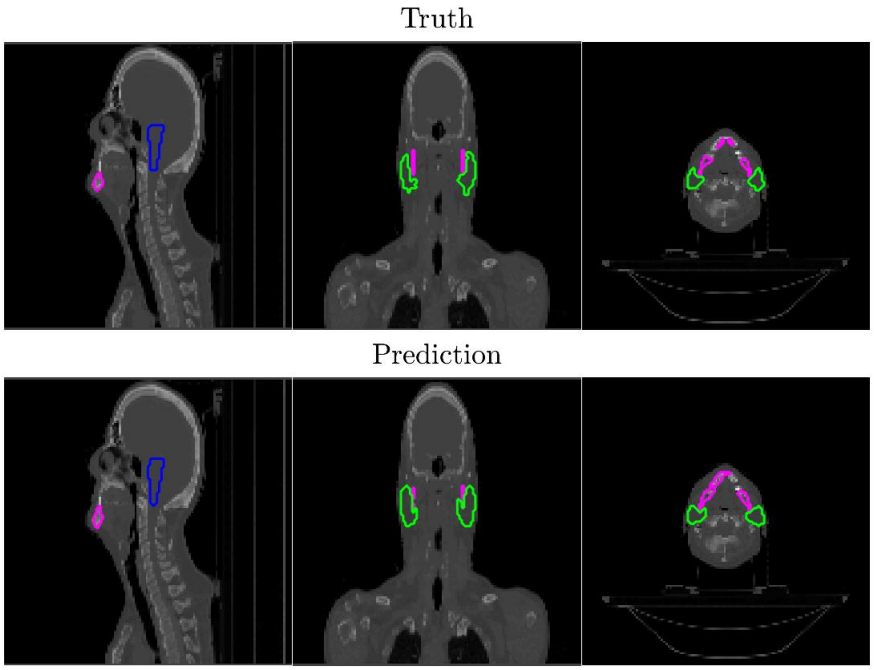


(e)

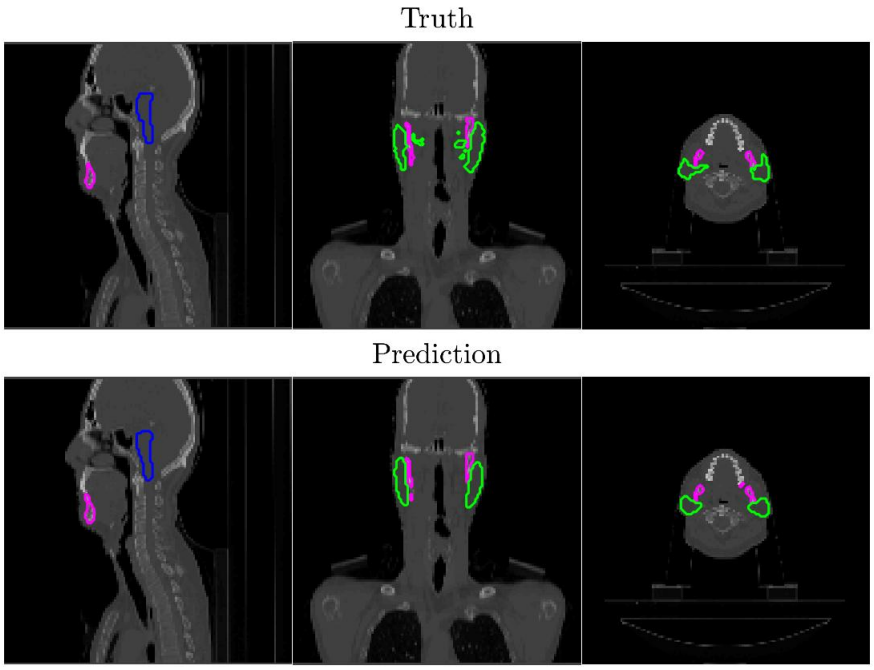
Figure 4.11: Subfigures 4.11a-4.11e display 3D models of the parotid glands comparing the ground truth to the predicted segmentation for the entire test set, with corresponding Dice values of: 0.7901, 0.7937, 0.835, 0.772 and 0.785. The *Truth+Prediction* is constructed by overlaying a transparent version of the prediction on top of the ground truth. The network was trained using reflection, distortion and rotation augmentation.

4.4 Segmented Model

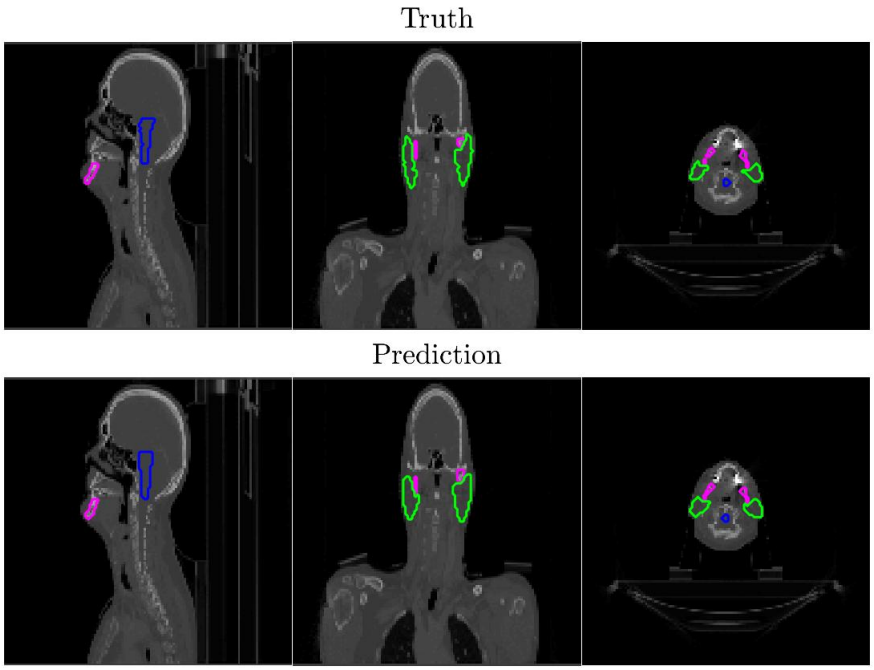
Figure 4.12 presents the final segmentation of all three organs for all examples in the test set. Note that in all subfigures, there are small but apparent differences between the ground truth and the predicted segmentation, especially so for the parotid glands.



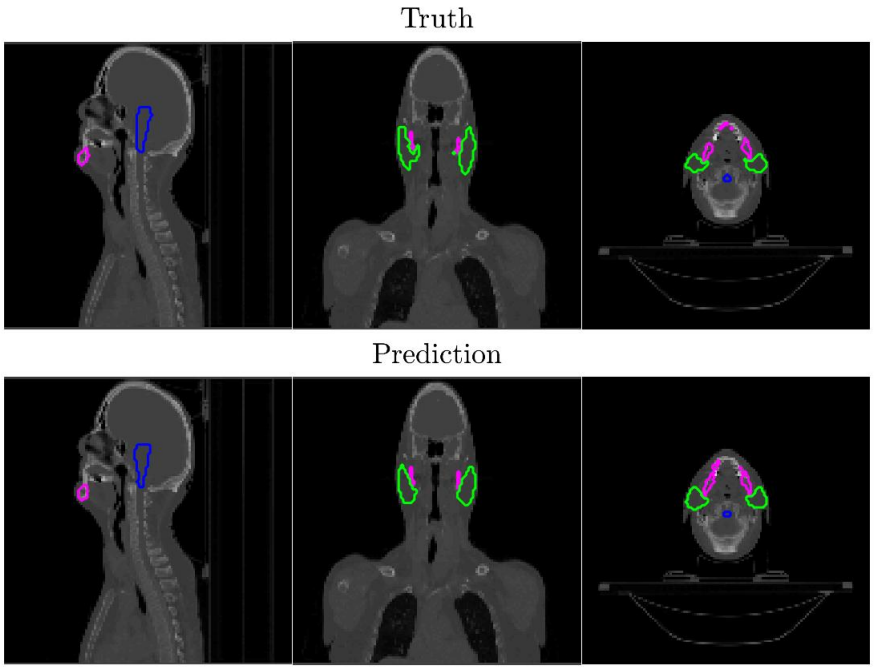
(a)



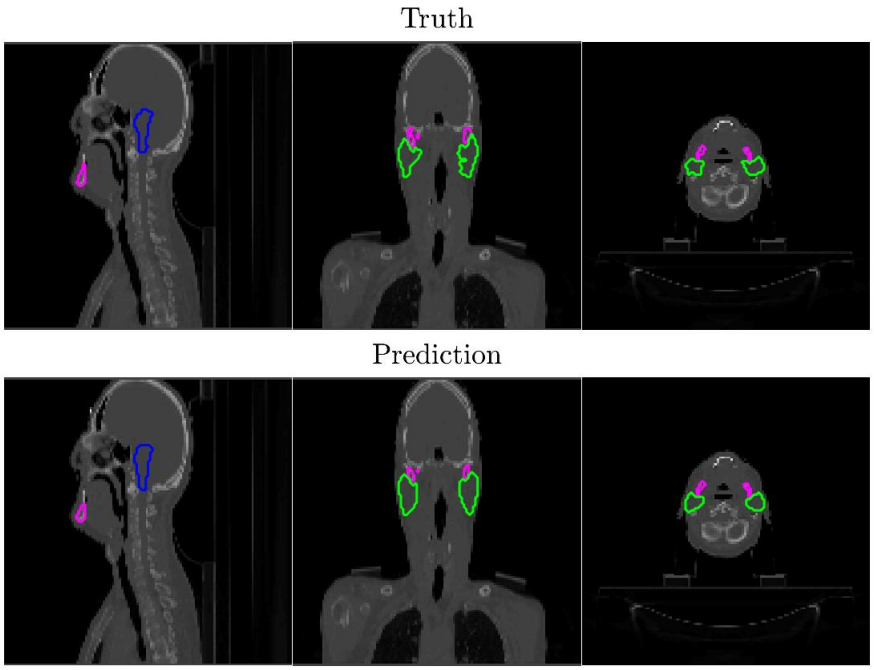
(b)



(c)



(d)



(e)

Figure 4.12: Subfigures 4.12a-4.12e present the median plane of the organ structures in all 3 dimensions comparing the ground truth to the segmentation for the entire test set, with corresponding Dice values of: 0.8498, 0.8638, 0.8674, 0.8405 and 0.8575. The mandible is marked in magenta, the brainstem in blue and the parotid glands in green. The network was trained using reflection, distortion and rotation augmentation for all organs.

5
Discussion

During the training process, one general observation about the loss convergence was that the main factor seemed to be the number of training examples per epoch. This is especially apparent given that the fastest convergence in terms of epochs occurred for training in patches of a higher resolution image. This is of course to be expected since when the network is exposed to more data in each epoch meaning that it naturally learns more with every epoch. The faster convergence in terms of epochs does, however, not mean that the total training time will be reduced which can be observed in Table 4.1. This is also reasonable since the network needs to be adjusted to more data. The difference does not, however, appear to be in direct relation to how much the data set has increased in size, which is promising as CNN:s are known to generate higher performance when exposed to more data. Consequently, the training time for a case with more available data is not expected to be proportionately higher which is advantageous.
In the beginning of Chapter 4, it is noted that the model trained with the corresponding loss convergence as in Subfigure 4.1c did not perform as well as a model trained according to the loss decay in Subfigure 4.1d. This is not necessarily intuitive, since the validation loss in Subfigure 4.1c is actually lower than the training loss, which indicates that the model has a higher performance on the validation set than the training set. This could be explained by that the network in this case is trained to generalize the segmentation problem more than what is required. Therefore, the network might make assumptions regarding the validation set which are not accurate. With this reasoning, it is possible that data augmentation might not improve the performance of the network when it is trained using the higher resolution patches. This, given that it is clear from Subfigure 4.1b that the training and validation loss coincide even without augmentation.
The following sections present a further discussion regarding the results obtained for each organ individually. Thereafter future prospects of the project will be examined followed by final conclusions drawn on the basis of this thesis.

5.1 The Mandible

Observing the bar charts presented in Figure 4.2, it is clear that the different methods examined during the ablation study perform similarly on the validation set with respect to all parameters. The patchwise training even results in a lower maximum Hausdorff index compared to the other methods, as can be seen in Subfigure 4.2e. This naturally indicates that the network has learned the segmentation of the val-

idation set next to equally well for the different cases. This is, however, not the case when it comes to segmentation of the test set. What is especially clear here is that the patchwise segmentation has a significantly lower performance compared to the other methods, including the teams of the MICCAI challenge, in respect to all evaluation metrics.
There is no great difference in between the different augmentation methods, but overall the case using all three augmentation methods is deemed to have the highest performance. This option having the highest mean Dice value and generally second to lowest, if not lowest, for the remaining parameters. Given that all three of the different augmentation methods lowered the convergence level of the validation loss, it is found reasonable that the use of a combination of all three methods would generate a higher performance. The fact that the different augmentations improve the performance also indicates that the chosen methods are feasible in generalizing the difference between different data sets. It is also to be expected that the difference in model performance is more apparent for segmentation of the test set of previously unseen data since this is a more clear indication of the generalization capabilities of the trained network. Comparing the fully augmented case to the top three teams from the MICCAI 2015 challenge, the performance in terms of obtained Dice values is similar as the VU and IM, for both the validation and test set, and somewhat higher than the UB team. However, apart from the patch-trained model, it is clear that this deep learning approach has a generally higher performance in terms of the different surface indices.
When comparing the mandible prediction with the highest Dice value to that with the lowest value obtained from the model trained with image patches the difference is grave, see Figure 4.3. The segmentation with the highest Dice coefficient, Subfigure 4.3a, is very similar to the corresponding ground truth, and of noticeably higher resolution compared to using the alternative data processing seen in Figure 4.5. However, the example with the lowest coefficient, Subfigure 4.3b, misses a large chunk of bone in the middle of the segmentation. This artifact is probably rooted in the fact that the network is trained using patches where sometimes the organ is divided in between different patches, which negatively affects the models ability to present a fully connected organ segmentation. This example is probably the reason why there is such a great difference on performance for the test set in respect to the evaluation metrics. Given that there is only 5 examples in the test set, one inadequate segmentation has great impact on the statistics for the entire set.
In general, it is clear from the 2D and 3D representations found in Figures 4.4 and 4.5 that the best final model, trained using all three augmentation methods, seem to generalize the shape of the mandible well. This is especially clear when observing the 3D representations, where both the ground truth and the prediction are so similar that they both appear to be adequate models of a mandible. With that meaning that given that the ground truth is manually segmented, the human factor makes it not fully reliable. The predicted segmentation could thus just as well be the more accurate, but this is of course difficult to validate.
The fact that the images are significantly resized to a dimension suitable for the network introduces artifacts on the finer organ structures, which for instance is clear in Subfigures 4.5b and 4.5c, where the ground truth no longer is fully connected.

This is problematic since the connectedness of the organ is an important property for the network to learn. The possibility of the organ not being fully connected appear to have been learned by the network, given that the prediction is not fully connected in Subfigure 4.5b. It is also possible that the difference in image resolution affects the comparability of these segmentations to those obtained in MICCAI 2015 since this somewhat changes the problem definition.
One reason as to why the network generated such accurate predictions of the mandible might be that there appear to quite a low variation in the ground truth in between the different data sets. This, based on the fact that the loss curves for the training and validation sets are so close even without the use of data augmentation, see Figure 3.2a. Observing the 3D representations of the test set presented in Figure 4.5, it is also clear that this organ has a characteristic appearance which is similar for all data sets. This could also be the reason as to why the top three teams from the MICCAI challenge also had a relatively high performance for this particular organ. Another reason might be that the fact that the mandible consist of bone, meaning that it is depicted in high contrast to the surrounding tissues given that the images are CT based.

5.2 The Brainstem

When observing the performance of the trained models on the test set in Figure 4.6, it is once again apparent that the augmented case performs consistently superior to the case with no augmentation. This option also competes well with the performance of the MICCAI teams, where the winner, IM, only performs slightly better in terms of Dice index but not in terms of the surface parameters. However, when observing the validation scores, there is a great difference in performance compared to the mandible evaluation. Since the augmented case missed predicting 2 out of 10 organs it had a grave impact on the mean Dice coefficient resulting in a lower overall performance compared to the case with no augmentation. It is also worth noting that the performance of the model trained without augmentation is slightly lower than all MICCAI teams for the validation set.
The mean measure is in general highly sensitive to outliers and given that the top 95 percentile is higher for the augmented case compared to the original, and it also performs better in terms of the surface indices where the missed predictions are excluded, there would not be such a significant difference if this measure were to be used instead. It is still highly problematic that the model missed the prediction of such a high percentage of the validation set. Given that this model still had a higher performance of the test set, it might be that in the process generalizing the segmentation problem further, it over-generalized in a manner unsuitable for the two missed cases. Since the evaluation of performance on a previously unseen test set is considered to be a more accurate measure of the actual model performance, the network trained using data augmentation is also in this case considered to be the premier choice.
Observing the 3D representations of the test set segmentation in Figure 4.8, it is clear that the network appears to have learned the general shape of the organ. There is, however, a more clear difference between the the ground truth and the prediction

compared to the mandible case, which also is reflected in the obtained values of the evaluation metrics. The general shape of this organ appears to vary more across the test set compared to the mandible case, which also is represented in the fact that there is a greater difference in between the validation and training loss convergence plateaus. When observing the 2D representation in Figure 4.7 there is one other apparent difference: lack of organ contrast. While the mandible was distinguished from its surroundings given that it was white, there is practically no visible difference in gray scale separating the brainstem from its surroundings.
Further observing the two examples in Figure 4.7, the organ positioning appears to be fairly accurate compared to the ground truth. This is probably due to the fact that the brainstem has a position which is quite consistent in terms of placement in relation to the skull bone and spinal cord. The lack of organ contrast does nonetheless hinder the model from distinguishing natural asymmetries in between the patients. This is especially clear when observing Subfigure 4.7b, where the organ is predicted to be completely symmetric for the two rightmost planes which is not the case in the corresponding ground truth. This could be an artifact from the reflection augmentation across the human symmetry axis. It is still not necessarily an erroneous generalization since this asymmetry does not appear to be consistently towards one lateral side, which is clear comparing the ground truths in Subfigures 4.7a and 4.7b.

5.3 The Parotid Glands

When observing the evaluation metrics for the predicted segmentation of the parotid glands presented in Figure 4.9, a similar analysis can be made as for the brainstem. There is once again a clearly higher performance of the model trained using data augmentation compared to the case with no augmentation when observing the test set performance. However, also in this case, the same is not true for the performance on the validation set. The fact that 2 out of 10 organs were missed by the model trained using data augmentation means that this once again introduced an instability which might be rooted in unsuitable generalization of the problem for the two missed organs. Also here, the performance for the surface indices, where the missed organs were excluded, indicated a higher quality of the segmentation which the model trained using augmentation actually do manage to segment. Given that the model trained using augmentation has a higher overall performance on the test set, it is also for this organ considered to be the superior choice.
Once again the performance of this model is similar to the MICCAI teams in terms of Dice index, although slightly lower than the IM team, but the 95 percentile HD and CM is consistently lower for both deep learning options. As for the maximum Hausdorff difference, there is a shift in performance and all three teams have overall lower scores. This indicates that the predicted segmentation of the parotid glands is somewhat poorer compared to the MICCAI teams in relation to the other organs. When observing the 2D and 3D representations in Figures 4.10 and 4.11 it is clear that a general shape and position if the glands is learned. Also, as for the mandible, finer structures appear to have lead to the ground truth being not fully connected after the resizing process. In this case, there is no indication that this quality is

learned by the network. It is possible that this can be explained by a larger organ shape variation and lack of contrast.
In Figure 4.11 there seems to be a significantly larger variation in organ shape compared to the other two organs. Also, as for the brainstem, Figure 4.10 indicates that even though there is a consistent positioning of the glands close to the mandible, there is next to no visible difference in contrast separating the organ from its surroundings. This, in combination with the large shape variation could be the main reason as to why these glands have been significantly more difficult for the trained models to segment in comparison to the mandible and brainstem.

5.4 Segmented Model

The overall performance of the predicted segmentation obtained by the use of this deep learning approach can be more clearly evaluated observing the images of the final combined prediction presented in Figure 4.12. From these images conclusions can be drawn that this method performs well when it comes to organ positioning and appearance but it lacks the finer details of the parotid glands. This is assumed to be rooted in the fact that their shape appears to vary more than the other organs across the test set with low image contrast separating them from their surroundings.
One general trend observed when comparing the segmentation performance in terms of the evaluation metrics to those obtained by the MICCAI teams was that the performance in terms of the Dice coefficient was similar but in respect to the surface indices the deep learning approach was superior in next to all cases. The fact the images which are segmented are resized to a dimension different from that used in the MICCAI challenge might be a factor since this somewhat changes the problem definition. Also, the network requires all training images to be of the same size which resulted in all images being re-scaled by different rate in the z-dimension which varied significantly across the set. This affected the overall anatomical accuracy of the patient representation. The surface indices are, however, evaluated in [mm], taking the new voxel spacing of the resized image into account, which means that the values should be comparable.
During the ablation study, it became clear that even though training the networks with patches of an image of higher resolution resulted in a more detailed prediction, this option was deemed more unreliable given that it predicted an organ which was completely detached in one example in the test set. This resulted in the conclusion that training on an image of lower resolution resized to fit the dimension for the network input was a more reliable option. When comparing the performance of models trained with the different augmentation methods implemented separately and combined to the case with no augmentation the augmented case showed the highest performance for both the validation and the test set. This resulted in this augmentation method being chosen for the two remaining organs
The test set performance of the trained models indicated that the fully augmented option generated a notably higher overall performance compared to the case with no augmentation for both the brainstem and the parotid glands. This was however not the case for the validation set were 2 organs out of 10 was missed for both organs for the network trained with augmentation. This shows that even though employing

augmentation increased the trained models ability to generalize the organ propertied for the test set, this generalization might not have been suitable for all examples of the validation set. It is therefore possible that this might not be an issue if the augmentation factors used were to be further optimized. The segmentation of both the brainstem and the parotid glands were significantly worse than for the mandible. It is also worth noting that there data sets available for the mandible was 8 fewer than the remaining organs, which makes this difference even more substantial. This behaviour is deemed to depend on both a larger shape variation in between different patients combined with a lack of contrast compared to the background.

6
Conclusion

In this project an automated 3D segmentation method was constructed to segment a set of CT images through the means of deep learning. Even through the final goal of this program is for it so serve as a step in the treatment planning process for antenna based local hyperthermia treatment of head and neck cancer, the final product is not quite there yet at this stage. Given that the data with a manual segmentation of the target structures was not provided in time for this project, the network could not be trained for these particular organs. Although, one of the segmented organs, the brainstem, is also an organ of interest for hyperthermia treatment. However, the use of different data inhibits the comparability to the current state-of-the-art.
The network was trained using organs of interest for radiology treatment planning, which actually makes the final program more suitable for this case at this stage. Comparing the performance of this method to other methods implemented by the use of the same data it is clear that this approach could be deemed to perform just as well as the winning team, which was model based. This conclusion is drawn taking into consideration that surface indices were overall lower and the Dice coefficient similar. This indicates that this method might perform higher than the current state-of-the-art for tissue segmentation in hyperthermia treatment planning. Especially so, since they use an atlas based approach as the third best team in the MICCAI challenge.
Even though both these methods are atlas based, there are of course variations in performance between different implementations as well. Just because this deep learning approach performs somewhat higher than an atlas based approach on the same issue, it does not necessarily mean that this would be the case compared to a different implementation on a different issue. However, based on the obtained results from this study, it could be advantageous to attempt to use this method for hyperthermia purposes since there is a possibility that this approach might perform higher than the current state-of-the-art. Finally, it is also worth noting that the current-state-of-the-art takes 3h/patient, which roughly can be translated to 16 min/organ. Given that a trained network only took around 2 s/organ, this approach has the prospect of greatly reducing the segmentation time which is one of the main factors as to why an automated segmentation approach is required in the first place.

6.1 Future Prospects

There are several options which could be examined to be able to make more well grounded conclusions and to further develop this project. For one, it would of course be interesting to train the network to predict segmentation of the remaining three structures from the MICCAI 2015 challenge and to also in these cases compare the results to those obtained by the top three teams.
It could also be advantageous to run experiments to further optimize the learning rate and learning rate decay to achieve a more consistent and reliable training convergence. Furthermore, it would be interesting to try a higher variation of distortion and augmentation factors to be able to find a more optimal choice for the training of each organ. It could also be interesting to try all different combinations of augmentation methods, meaning to also evaluate performance of training with only reflection and distortion versus using distortion and rotation compared also to rotation and reflection. It is perhaps so that the case where all augmentations are used together would still be the superior choice, however these trials would serve as a more solid basis for this conclusion.
Another option which could be explored to lower the overall training and prediction time is to train the network with multi-labeled data sets instead of binary organs. One down side to this is the fact that the data is not fully annotated. This would have to be taken into account, either by ignoring regions probable for the location of the missing organs of that data set or by excluding the data sets with missing organs. Either choice has their disadvantages. The first, given the risk of not fully removing the missing organ region and thus training the network for a structure which sometimes is seen as background and sometimes as foreground. The second option is also not ideal given that the training of CNN:s in general rely greatly on the amount of available data. However, these missing data sets could be somewhat compensated for by the use of data augmentation. The missing data would of course generate even more options if they also were to be augmented, but this might still be the preferred option.
To attempt to counteract the artifacts occurring when the image is resized to fit the network, it could be advantageous to look into adjusting the network to process larger image blocks. This would mean higher resolution of the final predicted segmentation without training with image patches, which showed to introduce a risk of producing disconnected organs. Unfortunately, a consequence would be a higher strain on the GPU which might mean, depending on how much the training image is enlarged, that the batch size would have to be reduced from 2 to 1. This could possible affect the performance of the final model. The possibility of simultaneous training on multiple strong GPU:s would eliminate this risk. However, this has not been possible in this project due to multi-threading issues related to the data storage. If this problem were to be solved it could even be possible to increase the batch size further.
Generally when comparing the predictions of the three evaluated organs it was clear that the mandible segmentation was superior to the two remaining structures. In the sections above it was discussed that one reason for this might be the lack of image contrast separating these organs, the brainstem ans parotid glands, from

their surroundings. This is, as previously mentioned, true for all soft tissues in CT images. However, this is not true for corresponding MR images, which are known for their high soft tissue contrast. Given that the organs of interest for segmentation in hyperhermia purposes are all of soft tissue it is possible that this would affect the final segmentation performance.
If the network on the other hand were to be trained using MR images instead this could possibly improve the results. One problem in this case would be the difficulty to thereafter put these segmented organs in relation to surrounding bone structures, which are not as well defined in MR images. These are, in addition to muscle and lungs, currently segmented by thresholding the CT contrast values, which would not be possible if MR images were to be used. One possible solution to this problem could be to segment soft tissues in MR images and bone, muscle and lungs using CT with a final image registration to combine the two. Given that these images are generated using different machines, this might be difficult since it could be a too a difference in patient positioning in the machine. Also, the difference in image contrast between the two makes it more difficult to find common features. Another option could be to train the network using MR images which also include a manual segmentation of the bone, muscle and lungs. Even if these have low contrast in MR, the same is true for soft tissues in CT images and this did not mean that they could not be predicted. The soft tissue structures of interest are of a higher quantity and it is possible that the bone, muscle and lungs vary less in shape in between patients in comparison.
In order to be able to make a fair evaluation whether this approach would be suitable to use for tissue segmentation in hyperthermia purposes is would naturally be necessary to use data which is manually segmented for the same purpose. If this data were to be provided by the creators of the current state-of-the-art, it would of course also be highly advantageous given that it would result in high comparability to their method. If one were to speculate the outcome of such a comparison based on the the results of this project it is possible that this method would be the superior choice. In their approach, they also discuss the advantages of combining CT and MRI images in order to generate a higher performance. For the mandible, which were or high contrast in the CT images, the deep learning approach outperformed the corresponding atlas based approach in respect to all quality indicators used. It can therefore be reasoned that this would be the case using MR images with high soft tissue contrast. This, in combination with it being computationally much faster, is a strong basis for choosing this approach. However, what might be problematic is the data limitations. This deep learning approach uses images of a significantly lower resolution compared to what it used in the MICCAI challenge. In order to be able to achieve a segmentation of the same high resolution as the current state-of-the-art for hypethermia treatment planning one would need a computer with a large RAM memory and the ability to combine several strong GPU:s for the training process.

Bibliography

[1] P. F. Raudaschl, P. Zaffino, G. C. Sharp, S. M. F., A. Chen, B. M. Dawant, and F. Jung, "Evaluation of segmentation methods on head and neck ct: Auto-segmentation challenge 2015", *Medical Physics*, vol. 44, no. 5, pp. 2020–2036, May 2017.

[2] B. Gupta, N. W. Johnson, and N. Kumar, "Global epidemiology of head and neck cancers: A continuing challenge", *Oncology*, vol. 91, no. 1, pp. 13–23, Jun. 2016.

[3] U. for International Cancer Control UICC. (2014). 2014 review of cancer medicines on the who list of essential medicines, [Online]. Available: `http://www.who.int/selection_medicines/committees/expert/20/applications/HeadNeck.pdf?ua%5C%20&hx003D;1` (visited on 06/05/2018).

[4] N. Vigneswaran and M. D.Williams, "Epidemiologic trends in head and neck cancer and aids in diagnosis", *Oral and Maxillofacial Surgery Clinics of North America*, vol. 26, no. 2, pp. 123–141, May 2014.

[5] S. Gao, M. Zheng, X. Ren, Y. Tang, and X. Liang, "Local hyperthermia in head and neck cancer: Mechanism, application and advance", *Oncotarget*, vol. 7, no. 35, pp. 57 367–57 378, Jun. 2016.

[6] V. Fortunati, "Automatic patient modeling for hyperthermia treatment planning of head and neck cancer", PhD thesis, Erasmus MC - University Medical Center Rotterdam, 2015.

[7] R. Verhaart and G. v. Rhoon, *Patient Modeling for Simulation Guided Head and Neck Hyperthermia*. Erasmus University Rotterdam, 2016, ISBN: 9789462955028. [Online]. Available: `https://books.google.se/books?id=5FQ%5C_nQAACAAJ`.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, `http://www.deeplearningbook.org`.

[9] B. Blaus. (2016). Brain anatomy (sagittal), [Online]. Available: `https://commons.wikimedia.org/wiki/File:Brain_Anatomy_(Sagittal).png`.

[10] O. College. (2013). Illustration from anatomy & physiology, connexions web site, [Online]. Available: `https://commons.wikimedia.org/wiki/File:1801_The_Endocrine_System.jpg`.

[11] N. E. Institute. (2010). Anatomy of the eye, [Online]. Available: `https://www.flickr.com/photos/nationaleyeinstitute/7544655864`.

[12] C. OpenStax. (2016). Openstax anatomy and physiology, [Online]. Available: `https://commons.wikimedia.org/wiki/File:1204_Optic_Nerve_vs_Optic_Tract.jpg`.

[13] ——, (2016). Biology, [Online]. Available: `https://commons.wikimedia.org/wiki/File:Figure_34_01_08ab.jpg`.

[14] B. Blaus. (2013). Facial bones, [Online]. Available: https://commons.wikimedia.org/wiki/File:Blausen_0393_FacialBones_01.png.

[15] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool", *BMC Medical Imaging*, vol. 15, no. 1, p. 29, Aug. 2015, ISSN: 1471-2342. DOI: 10.1186/s12880-015-0068-x. [Online]. Available: https://doi.org/10.1186/s12880-015-0068-x.

[16] N. Sharma and L. M. Aggarwal, "Automated medical image segmentation techniques", *Journal of Medical Physics*, vol. 35, no. 1, pp. 3–14, Aug. 2010.

[17] T. Heimann and H. Delingette, "Model-based segmentation", in *Biomedical Image Processing*, T. M. Deserno, Ed., Springer, 2011, pp. 279–303. DOI: 10.1007/978-3-642-15816-2_11. [Online]. Available: https://hal.inria.fr/inria-00616063.

[18] X. Gao, Y. Su, X. Li, and D. Tao, "A review of active appearance models", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 2, pp. 145–158, Mar. 2010, ISSN: 1094-6977. DOI: 10.1109/TSMCC.2009.2035631.

[19] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey", *CoRR*, vol. abs/1412.3421, 2014. arXiv: 1412.3421. [Online]. Available: http://arxiv.org/abs/1412.3421.

[20] V. Duay, N. Houhou, and J. P. Thiran, "Atlas-based segmentation of medical images locally constrained by level sets", in *IEEE International Conference on Image Processing 2005*, vol. 2, Sep. 2005. DOI: 10.1109/ICIP.2005.1530298.

[21] O. Enqvist, *Lecture notes in image analysis*, Jan. 2017.

[22] H. Kalini, "Atlas-based image segmentation: A survey", Apr. 2018.

[23] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation", *Annual Review of Biomedical Engineering*, vol. 2, no. 1, pp. 315–337, 2000, PMID: 11701515. DOI: 10.1146/annurev.bioeng.2.1.315. eprint: https://doi.org/10.1146/annurev.bioeng.2.1.315. [Online]. Available: https://doi.org/10.1146/annurev.bioeng.2.1.315.

[24] F. Dietzel, "Vascular perfusion in cancer therapy. recent results in cancer research (fortschritte der krebsforschung / progrès dans les recherches sur le cancer)", in, K. Schwemmle and K. Aigner, Eds. Heidelberg: Springer-Verlag Berlin, 1983, vol. 86, ch. Basic Principles in Hyperthermic Tumor Therapy, pp. 177–190.

[25] J. Prince and J. Links, *Medical Imaging Signals and Systems*. Pearson, 2014, ISBN: 9780132145183.

[26] M. S. Fasihi and W. B. Mikhael, "Overview of current biomedical image segmentation methods", *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 803–808, Dec. 2016.

[27] I. Blausen Medical Communications. (2013). Cat scan, [Online]. Available: https://commons.wikimedia.org/wiki/File:Blausen_0205_CATScan_01.png.

[28] B. Zitova and J. Flusser, "Image registration methods: A survey", *Image and Vision Computing*, vol. 21, pp. 977–1000, Jun. 2003.

[29] J. M. Fitzpatrick, D. L. G. Hill, and C. R. Maurer Jr., "Handbook of medical imaging: Medical image processing and analysis", in. SPIE Press - The Interna-

tional Society for Optical Engineering, 2004, vol. 2, ch. 8 - Image Registration, pp. 447–513.

[30] S. Yuheng and Y. Hao, "Image segmentation algorithms overview", *CoRR*, vol. abs/1707.02051, 2017. arXiv: 1707.02051. [Online]. Available: http://arxiv.org/abs/1707.02051.

[31] M. A. Nielsen, *Neural Networks and Deep Learning*. Determination Press, 2015.

[32] W. Commons. (2015). Max pooling with 2x2 filter and stride = 2, [Online]. Available: https://commons.wikimedia.org/wiki/File:Max_pooling.png.

[33] S. Shi and X. Chu, "Speeding up convolutional neural networks by exploiting the sparsity of rectifier units", *CoRR*, vol. abs/1704.07724, 2017. arXiv: 1704.07724. [Online]. Available: http://arxiv.org/abs/1704.07724.

[34] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 challenge", *CoRR*, vol. abs/1802.10508, 2018. arXiv: 1802.10508. [Online]. Available: http://arxiv.org/abs/1802.10508.

[35] S. Ruder, "An overview of gradient descent optimization algorithms", *CoRR*, vol. abs/1609.04747, 2016. arXiv: 1609.04747. [Online]. Available: http://arxiv.org/abs/1609.04747.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *CoRR*, vol. abs/1412.6980, 2014. arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980.

[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", *CoRR*, vol. abs/1502.03167, 2015. arXiv: 1502.03167. [Online]. Available: http://arxiv.org/abs/1502.03167.

[38] P. M. Radiuk, "Impact of training set batch size on the performance of convolutional neural networks for diverse datasets", vol. 20, Dec. 2017.

[39] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization", *CoRR*, vol. abs/1607.08022, 2016. arXiv: 1607.08022. [Online]. Available: http://arxiv.org/abs/1607.08022.

[40] L. Taylor and G. Nitschke, "Improving deep learning using generic data augmentation", *CoRR*, vol. abs/1708.06020, 2017. arXiv: 1708.06020. [Online]. Available: http://arxiv.org/abs/1708.06020.

[41] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning", *CoRR*, vol. abs/1712.04621, 2017. arXiv: 1712.04621. [Online]. Available: http://arxiv.org/abs/1712.04621.

[42] G. J. S. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis", *CoRR*, vol. abs/1702.05747, 2017. arXiv: 1702.05747. [Online]. Available: http://arxiv.org/abs/1702.05747.

[43] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation", *CoRR*, vol. abs/1606.06650, 2016. arXiv: 1606.06650. [Online]. Available: http://arxiv.org/abs/1606.06650.