Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Physics

Kaja Slåtsve Øvrelid

Artificial intelligence-based automatic segmentation for breast cancer radiotherapy

Master's thesis in Applied Physics and Mathematics
Supervisor: Sigrun Saur Almberg, St. Olavs Hospital
June 2020

NTNU
Norwegian University of
Science and Technology

ST. OLAVS HOSPITAL
TRONDHEIM UNIVERSITY HOSPITAL

Kaja Slåtsve Øvrelid

Artificial intelligence-based automatic segmentation for breast cancer radiotherapy

Master's thesis in Applied Physics and Mathematics
Supervisor: Sigrun Saur Almberg, St. Olavs Hospital
June 2020

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Physics

NTNU
Norwegian University of
Science and Technology

Abstract

Background and purpose: Accurate segmentation of target volumes and organs at risk is critical for the patient treatment outcome in radiotherapy. Manual segmentation of structures is known as the largest uncertainty in the radiotherapy process. Auto-segmentation based on artificial intelligence (AI) may lead towards a faster and more consistent way of contouring. The aim of this study was to investigate two different methods using AI for automatic segmentation of relevant structures for radiotherapy treatment planning of breast cancer patients. This included evaluating a deep learning (DL) thorax model, implemented in a commercial treatment planning system, and training and testing machine learning (ML) models, implemented in Python.

Materials and method: All patient data was from left-sided breast cancer patients previously treated with external photon beam radiotherapy at St. Olavs Hospital, using deep inspiration breath hold. The DL thorax model was evaluated quantitatively and clinically for 20 patients by generating segmentations for the heart, the lungs, the spinal cord, and the esophagus. For segmentation of the sternum, the left breast, and the heart, ML models using linear support vector classification were trained with 20 and 30 patients and evaluated quantitatively. The Dice similarity coefficient (DSC), percentile Hausdorff distances (HDs), and the average HD (AVD) were used for quantitative evaluation.

Results: The DL thorax model used on average 3 minutes on generating AI segmentations for one patient. The average DSC for the heart and lungs were 0,92 ± 0,02 and 0,97 ± 0,01, respectively; the average AVD for the heart and the lungs were 2,9 ± 1,1 mm and 0,9 ± 0,4 mm, respectively. In terms of clinical acceptability, the AI-generated segmentations passed in 42 % of the cases for the heart, 100 % of the cases for the lungs, 85 % of the cases for the spinal cord, and 70 % of the cases for the esophagus. The runtime for the ML models was on 30 seconds to 5 minutes. For the models trained with 30 patients, the average DSC for the sternum, the left breast, and the heart were 0,65 ± 0,06, 0,64 ± 0,10, and 0,66 ± 0,05, respectively; the average AVD for the sternum, the left breast, and the heart were 1,8 ± 0,6 mm, 2,3 ± 0,5 mm, and 2,4 ± 0,5 mm, respectively.

Conclusion: Regions of interest (ROIs) can easily be contoured with a DL thorax model for breast cancer patients. Along with high accuracy, a large majority of the segmentations were clinically acceptable, and many of the non-accepted segmentations required minor manual corrections. This implies that the model has the potential to improve both consistency and efficiency of segmentation in the clinic. The ML algorithm can easily be trained to contour ROIs for breast cancer patients; however, the ML models need further improvements in order to be clinically useful.

Sammendrag

Bakgrunn og formål: Nøyaktig inntegning av målvolum og risikoorganer er avgjørende for resultatet av pasientbehandling med stråleterapi. Manuell inntegning av strukturer er kjent som den største usikkerheten i stråleterapiprosessen. Automatisk segmentering ved bruk av kunstig intelligens (AI) kan gi en raskere og mer konsistent måte å tegne inn strukturer på. Målet med denne studien var å undersøke to forskjellige AI-metoder for automatisk segmentering av relevante strukturer for strålebehandling av brystkreftpasienter. Dette inkluderte å evaluere en dyp læring (DL)-thoraxmodell, implementert i et kommersielt doseplanleggingssystem, og å trene og teste maskinlæring (ML)-modeller, implementert i Python.

Materiale og metode: All pasientdata var fra venstresidig brystkreftpasienter som har blitt behandlet med ekstern stråleterapi med fotoner ved St. Olavs hospital, ved bruk av pustestyring. DL-thoraxmodellen ble evaluert kvantitativt og klinisk for 20 pasienter ved å generere inntegninger for hjertet, lungene, ryggmargen og spiserøret. For segmentering av brystbenet, venstre bryst og hjertet, ble ML-modeller som bruker lineær støttevektorklassifisering trent med 20 og 30 pasienter og evaluert kvantitativt. Dice score (DSC), Hausdorff-avstand (HD)-persentiler og gjennomsnittlig HD (AVD) ble brukt til kvantitativ evaluering.

Resultater: DL-thoraxmodellen brukte i gjennomsnitt 3 minutter på å generere AI-segmenteringer for én pasient. Gjennomsnittlig DSC for hjerte og lunger var henholdsvis 0,92 ± 0,02 og 0,97 ± 0,01; gjennomsnittlig AVD for hjerte og lunger var henholdsvis 2,9 ± 1,1 mm og 0,9 ± 0,4 mm. I den kliniske analysen passerte de AI-genererte segmenteringene i 42 % av tilfellene for hjertet, 100 % av tilfellene for lungene, 85 % av tilfellene for ryggmargen og 70 % av tilfellene for spiserøret. Kjøretiden for ML-modellene var på 30 sekunder til 5 minutter. For modellene trent med 30 pasienter, var gjennomsnittlig DSC for brystbenet, venstre bryst og hjertet henholdsvis 0,65 ± 0,06, 0,64 ± 0,10 og 0,66 ± 0,05; gjennomsnittlig AVD for brystbenet, venstre bryst og hjertet var henholdsvis 1,8 ± 0,6 mm, 2,3 ± 0,5 mm og 2,4 ± 0,5 mm.

Konklusjon: Strukturer kan enkelt tegnes inn med en DL-thoraxmodell for brystkreftpasienter. Sammen med høy nøyaktighet var et stort flertall av segmenteringene klinisk aksepterte, og mange av de ikke-aksepterte segmenteringene krevde kun mindre manuelle korreksjoner. Dette innebærer at modellen har et potensiale til å forbedre både konsistensen og effektiviteten av segmentering i klinisk praksis. ML-algoritmen kan lett trenes til å tegne inn strukturer for brystkreftpasienter; ML-modellene må imidlertid forbedres ytterligere før de kan brukes i klinisk praksis.

Preface

This master thesis is submitted as the conclusion of the master's degree program in Applied Physics and Mathematics at the Norwegian University of Science and Technology (NTNU). The presented work was performed during the spring semester of 2020 at the Department of Radiotherapy, Cancer Clinic at St. Olavs Hospital in Trondheim.

First of all, I would like to thank my supervisors Sigrun Saur Almberg and Kathrine Røe Redalen for involving me in such an interesting project. Sigrun has been especially helpful and encouraging during my work with this master thesis, along with answering all my questions and giving me valuable feedback in the writing process. She has made it possible for me to finish my master thesis despite limited access to the hospital due to the COVID-19 outbreak. Kathrine has included me in her research group, which has been both educational and very pleasant. Being a part of this group has motivated me throughout the semester. I would also like to thank PhD student Franziska Knuth for her helpful guidance on training the machine learning models and for providing me with thoroughly feedback. Also, a big thank you to oncologist Monika Eidem for the clinical evaluation of the segmentations.

Lastly, I would like to thank my friends and family and all my fellow students at Biophysics and medical technology with whom I have spent the last years.

Trondheim, 15-06-2020

Kaja Slåtsve Øvrelid

Contents

Abstract i
Sammendrag iii
Preface v
Contents vii
Abbreviations ix
1 Introduction 1
2 Theory 3
2.1 External beam radiotherapy 3
2.1.1 Radiotherapy workflow 3
2.1.2 The linear accelerator 4
2.1.3 Treatment techniques 5
2.2 Auto-segmentation methods 6
2.2.1 Traditional auto-segmentation 6
2.2.2 Artificial intelligence for auto-segmentation 7
2.3 Artificial intelligence 9
2.3.1 Machine learning 10
2.3.2 Deep learning 13
2.3.3 Artificial intelligence in radiation oncology 16
2.4 Breast cancer 18
2.4.1 Anatomy and physiology 18
2.4.2 Treatment modalities 18
3 Materials and method **21**
3.1 Evaluation of a DL thorax model 21
3.1.1 The model 21
3.1.2 Patient data 22
3.1.3 Evaluation 23
3.2 Training and testing of ML models 24
3.2.1 The algorithm 24
3.2.2 Patient data 24
3.2.3 Training and testing 25
3.3 Description of methods used for comparison 26
3.3.1 Quantitative evaluation 27
3.3.2 Clinical evaluation 28
3.3.3 Statistical analysis 29

4 Results **31**
4.1 Evaluation of a DL thorax model 31
4.1.1 Heart 31
4.1.2 Lungs 34
4.1.3 Spinal cord and esophagus 38
4.2 Training and testing of ML models 40
4.2.1 Sternum 41
4.2.2 Left breast 43
4.2.3 Heart 45
5 Discussion **47**
5.1 Evaluation of a DL thorax model 47
5.2 Training and testing of ML models 51
5.3 Metrics used for comparison 53
5.4 Management of patient data 56
5.5 Further work 58
6 Conclusion **61**
Bibliography **63**
A Python script for extracting dose values in RayStation **69**
B Python script for calculating quantitative metrics in RayStation **71**
C Normality test **73**

Abbreviations

3D-CRT Three-dimensional conformal radiotherapy. 5

AI Artificial intelligence. i, iii, 1

ANN Artificial neural network. 13

AVD Average Hausdorff distance. i, iii, 27

CNN Convolutional neural network. 1

CT Computed tomography. 1

DL Deep learning. i, iii, 1

DNN Deep neural network. 13

DSC Dice similarity coefficient. i, iii, 27

DVH Dose-volume histogram. 5

GPU Graphics processing unit. 21

HD Hausdorff distance. i, iii, 27

IMRT Intensity modulated radiotherapy. 5

LAD Left anterior descending coronary artery. 19

linac Linear accelerator. 3

ML Machine learning. i, iii, 1

MLC Multileaf collimator. 5

MR Magnetic resonance. 3

OAR Organ at risk. 1

PET Positron emission tomography. 3

QA Quality assurance. 3

ReLU Rectified linear unit. 15

ROI Region of interest. i, 1

SVC Support vector classification. 11

SVR Support vector regression. 11

TP False negative. 27

TP False positive. 27

TP True positive. 27

VMAT Volumetric modulated arc therapy. 5

1 Introduction

Radiotherapy is always a balance between destroying the cancer cells and minimizing damage to healthy tissue. For every patient that is to receive radiotherapy, a tailor-made treatment plan is generated. The treatment plan needs to be based on an accurate anatomical model of the patient. Target volumes and organs at risk (OARs) are traditionally manually contoured on a computed tomography (CT) scan by a physician. Segmentation of the tumor and OARs is known as the largest uncertainty in the process of radiotherapy, and accurate segmentation is critical for the patient treatment outcome [1, 2].

The segmentation quality and time spent on contouring strongly depend on the experience of the practitioner and complexity of the case [3, 4], a process that can take anywhere from 30 minutes to many hours. Semi-automatic methods for segmentation have been devised and are useful tools for speeding up the process. Nonetheless, the process of manually segmenting regions of interest (ROIs) is time-consuming and suffers from intra- and interobserver variability. With improved automatic tools, this process can be greatly simplified and lead towards a faster and more consistent way of contouring.

The automation of the radiotherapy planning process is both desirable and challenging. In the later years, there have been substantial technological developments in the field of artificial intelligence (AI), also in radiation oncology. Modern computer technology now enables the use of AI in radiotherapy planning, and auto-segmentation approaches using machine learning (ML) algorithms and deep learning (DL) algorithms based on convolutional neural networks (CNNs) have recently become clinically available [5, 2]. These methods can improve efficiency and consistency; with this comes a potential for better use of resources and improved quality of treatment planning [6, 7]. However, before clinical use, these methods need thorough evaluation, and clinically relevant contour evaluation remains challenging.

This master thesis was carried out to investigate two different AI methods for automatic segmentation of relevant structures for radiotherapy treatment planning of breast cancer patients. This process includes different aspects, and the specific aims of this thesis were to

1. Evaluate the performance of a previously trained DL thorax model in RayStation (RaySearch Laboratories AB, Stockholm, Sweden), in terms of accuracy and clinical applicability.
2. Train ML models for segmentation of structures relevant for breast cancer treatment and test them in terms of accuracy.

2 Theory

2.1 External beam radiotherapy

Radiotherapy utilizes ionizing radiation to treat cancer, either for cure or palliation. Radiotherapy is delivered most commonly by a medical linear accelerator (linac), where high-energy X-rays with energies of 6-15 MV or electron radiation with energies of 6-18 MeV are typically used. When using a linac, it is called external beam radiotherapy, because the radiation enters the patient from outside.

2.1.1 Radiotherapy workflow

The radiotherapy process can be divided into different stages: patient assessment, simulation, treatment planning and quality assurance (QA), treatment delivery and monitoring, and follow-up [8]. Figure 1 presents a typical radiotherapy workflow.

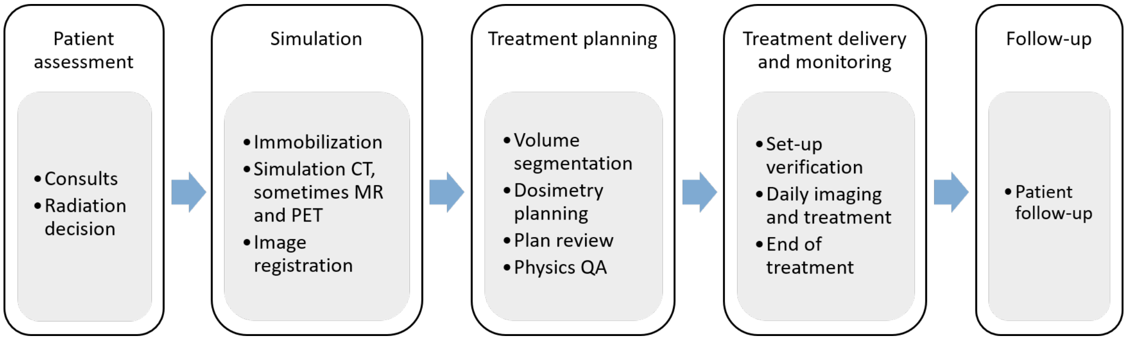


Figure 1: Radiotherapy workflow, from patient consult and assessment to follow-up.

The radiotherapy process begins at the first consultation, where the clinical situation is discussed and risks and benefits of treatment are considered. If it is decided to proceed with radiotherapy, a CT scan of the patient is taken. This requires careful positioning and immobilization of the patient as the treatment must be reproducible over many fractions. Further instructions include details about scan range, treatment site, and other specifics necessary to complete the procedure appropriately. When the CT simulation is completed and reviewed, the images are exported to a treatment planning system. If necessary, magnetic resonance (MR) imaging can provide additional information for soft tissue contouring, and positron emission tomography (PET) scanning can be used to identify the biological characteristics of the tumors. The full set of image data serves as a three-dimensional anatomical model of the patient, and the planning process starts with the segmentation of target volumes and OARs. The planning process continues by selecting an appropriate treatment technique, setting dosimetric goals for targets and normal tissues, and iteratively modifying different parameters until the planning goals have been achieved. This is always a compromise between destroying the cancer cells and minimizing damage to the normal cells. Finally, the plan is evaluated and approved [8]. Additionally, QA is embedded in each step of the process to ensure the safe delivery of radiotherapy. Likewise, the patient follow-up begins at the start of the treatment and continues after the end of the treatment.

2.1.2 The linear accelerator

The following section is based on [9] and [10]. Some details may be relevant for Elekta linacs only. The linac delivers high-energy X-rays or electrons to the region of a patient's tumor. The electron beam is useful for the treatment of superficial tumors down to about 5 cm depth, but for more deep-seated tumors, it is better to use several photon beams combined in a cross-fire. The linac is mounted on a drum structure, named the gantry, which can rotate through 360 degrees around the patient. The gantry enables the beam to be directed towards the patient from any direction. To ensure precise delivery of complex treatment plans, the accuracy of rotation must be less than 2 mm. A simplified illustration of the linac and its components are shown in Figure 2.

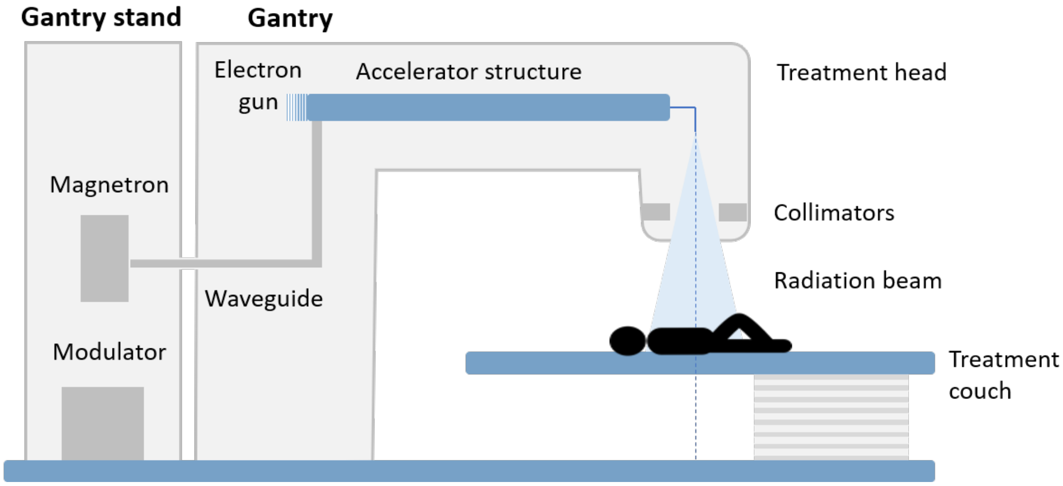


Figure 2: Sketch of a linac. The microwaves generated by the magnetron are guided into an accelerating waveguide, where they are used to accelerate electrons supplied from the electron gun. Further, the electrons are deflected by a magnet and directed towards the patient. Patients are treated either using the electrons directly or by creating bremsstrahlung photons.

The linac uses microwave technology to accelerate electrons in a part of the accelerator called the waveguide. The waveguide is a metal tube, which is fed with propagating radio frequency waves produced by the magnetron. The magnetron controls the power and frequency of these radiofrequency waves. This action is synchronized with the injection of electrons by the electron gun. The electrons are produced by heating a tungsten filament within the cathode, and the number of electrons injected is controlled by the temperature of the filament. Furthermore, the electrons must have the right phase relative to the radiofrequency waves in order to gain energy and be accelerated along the waveguide.

The waveguide contains a series of small metal irises that increase the wavelength of the microwaves. At the same time, the frequency is constant, accelerating pulses of electrons almost to the speed of light. Also, a vacuum is created to ensure that other particles do not embed the electron beam. The linac must produce a stable electron beam concentrated onto a small focal spot. Thus, the focussing and steering of the beam are controlled by modifying the current in different electromagnets. Two sets of focussing coils provide a static, axial magnetic field, which helps to limit the radius of the beam, whereas two sets of steering coils provide beam centering. The electrons are then deflected by bending magnets to be directed towards the patient.

Patients are treated either using the electrons directly or by creating bremsstrahlung photons. The latter is achieved by letting the electrons collide with a heavy metal target to produce high-energy X-rays. The high-energy X-rays are then shaped as they exit the machine, usually by a multileaf collimator (MLC) that is incorporated into the head of the machine.

2.1.3 Treatment techniques

There are different techniques for delivering external radiotherapy. Common techniques include three-dimensional conformal radiotherapy (3D-CRT), intensity-modulated radiotherapy (IMRT), and volumetric modulated arc therapy (VMAT). Most types of radiotherapy treatments use photons, and the mentioned techniques are therefore presented for treatment with photon beams.

3D-CRT uses several fields that are shaped by a MLC to conform the dose to the target volume while shielding normal tissues. In this way, a more uniform dose is delivered to the target volume and the dose received by the OARs is reduced. The 3D-CRT process involves forward-planning to create radiation dose distributions. In forward-planning, the number, direction, beam weighting, and shapes of the radiation beams are defined by the treatment planner [11]. A plan is commonly evaluated based on visual inspection of the dose distribution and dose-volume histogram (DVH)-data. This method is time consuming, and it is not possible to explore all options [12].

Instead, more conformal and complex dose distributions can be obtained with modern planning techniques. Modern treatment planning systems have implemented inverse planning algorithms. In inverse planning, the main focus is the final dose distribution and not how this dose distribution is accomplished; it starts with a description of the desired dose distribution and derives the beam shapes as a second step. This is accomplished by defining an objective function, which is an expression of how well the actual dose distribution compares to the requested dose distribution. The optimization algorithm bases its strategy on the objective function and choose parameters that make an improvement in the dose distribution. This is an iterative process where the goal is to minimize the objective function and find the global minimum [12].

Today, IMRT and VMAT are becoming routine for most treatment planning in the clinic [12]. IMRT allows for the creation of irregular-shaped radiation doses that conform to the tumor whilst simultaneously avoiding critical organs. In this technique, not only the shape but also the intensity profile, or the fluence, of each beam is modulated. This makes IMRT superior to the 3D-CRT technique. For IMRT, the dose-volume requirements must be explicitly expressed. This includes both dose to the target volume and acceptable dose limits for the OARs. Through a step-by-step process, the planning program searches for intensity distributions in the radiation fields that provide the best dose distribution and that meet all dose-volume requirements.

VMAT is an advanced form of IMRT that delivers the radiation dose continuously as the treatment machine rotates around the patient. With information about the linac, the treatment planning system calculates how the treatment device should rotate, how the MLC should move, and how the dose rate should vary. Unlike IMRT treatments, where the treatment machine make repeated stops and treat the tumor from a number of different angles, VMAT can deliver dose to the entire tumor in one single gantry rotation without any stops. This significantly reduces the average treatment time per fraction compared to IMRT [13].

2.2 Auto-segmentation methods

Segmentation of medical images aims to locate anatomic structures and contour their boundaries on a digital source. In radiotherapy, image segmentation is an important task routinely performed to identify the treatment target and the OARs that are to be avoided during irradiation. The ROIs are traditionally segmented manually by a physician, and the radiotherapy dose calculation is primarily done on CT scans. In some clinics, however, MR imaging is also being used more frequently [2]. Manual segmentation is still the standard routine for most clinics, although it is time consuming and prone to intra- and interobserver variations. Automated segmentation methods seek to decrease the time of segmentation and standardize the anatomical structure definition.

2.2.1 Traditional auto-segmentation

The development of auto-segmentation algorithms is related to how well algorithms utilize prior knowledge for new segmentation tasks. Traditional auto-segmentation approaches can be grouped as atlas-based segmentation and model-based segmentation, depending on the amount of historical patient and plan data used in the algorithms [5].

Atlas-based

Atlas-based segmentation methods generate a novel set of segmentations from a previously labeled, segmented reference image. The reference image is referred to as an atlas and contains information on locations and shapes of anatomical structures and the spatial relationships between them. For example, an atlas can be generated by manually segmenting a selected image or by integrating information from multiple segmented images. In single atlas-based segmentation, one reference image with segmented ROIs is used as a template for new segmentation tasks, while multi-atlas segmentation uses a number of atlases to compensate for variability between subjects [14].

Although many variations exist, the general approach is to map segmentations from a similar patient onto a novel patient using deformable image registration. The image is then segmented by mapping its coordinate space to that of the atlas, in an anatomically correct way, by finding the optimal transformation between the atlas and the new image. This process is known as the registration, and by mapping an image to an atlas, the label for each image voxel can be determined by looking up the structure at the corresponding location in the atlas under that mapping [14].

Model-based

Model-based segmentation techniques contours organs automatically using statistical shape or appearance models for different body sites. These models utilize a set of contoured images to recognize characteristic variations of shape or appearance of structures of interest. In this approach, an organ model is first positioned over the anatomical structure in the image set, and a deformable model algorithm then adapts the organ model to the boundaries of the anatomical structure. However, the limitation of specific shapes characterized by the statistical models makes this approach less flexible. Another limiting factor is the size and quality of the training data available [5].

2.2.2 Artificial intelligence for auto-segmentation

In the later years, there have been substantial technological developments in the field of AI, also in radiation oncology. Recent works in the field of medical image segmentation have used AI to automate the image segmentation task, and algorithms using ML and DL have recently become clinically available. ML- and DL-based segmentation can be used in auto-segmentation when larger amounts of contoured images are available for training. The algorithms can learn appropriate priors for structures by using an extensive patient database as input to train the segmentation model [5]. A more general description of AI, ML, and DL is given in chapter 2.3.

In order to achieve auto-segmentation using AI, one must first train a model. The workflow for training an AI model is shown in Figure 3. The model is trained on a representative dataset, which means CT scans with segmented structures from anonymized patient data for the appropriate diagnosis and treatment site. To get the most out of the available data, data augmentation may be used to artificially expand the size of the training dataset by creating modified versions of the images in the dataset. The augmentation includes image transformations, such as small random rotations and translations. This is performed during the training of the algorithm, meaning that the model is presented with slightly different versions of the images for each training iteration. The algorithm is optimized so that there is one for each ROI. After this, the trained model is completely anonymized and does not contain any image data from the training dataset [15].

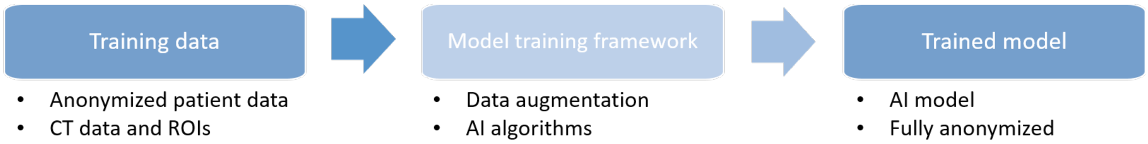


Figure 3: Process of training an AI model for organ segmentation. The image data is used to train the algorithm to produce the trained model. With ML, relevant features must be manually extracted from the input data and fed to the algorithm; with DL, the DNNs automatically extract relevant features.

When the model is applied to a new patient geometry, the input consists of the trained model and the new CT data, as illustrated in Figure 4. For ML models, unique features, such as shape or edges, must be identified, extracted, and given as input to the algorithm. DL algorithms do not require feature extraction and can be applied directly to the input data. The trained neural network can be thought of as a non-linear function taking a three-dimensional image as input and producing a labeled image as output. The CT image stack is pushed through the neural networks to predict ROIs. Finally, post-processing may be used to further improve the segmentation result.

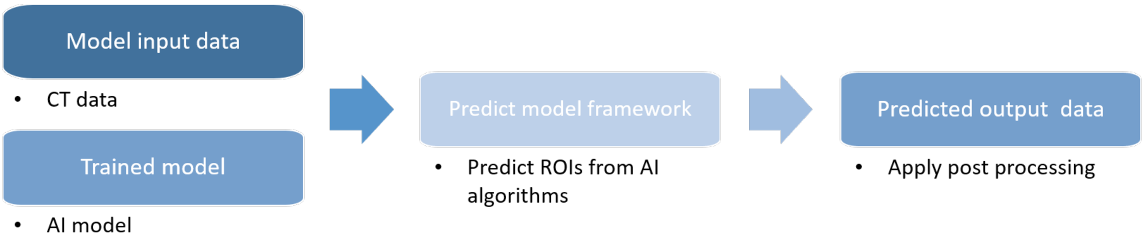


Figure 4: Process of applying an AI model for organ segmentation. The trained model takes the image data as input and outputs the labeled image. With ML, relevant features must be manually extracted from the input data and fed to the algorithm; with DL, the DNNs automatically extract relevant features.

Conventional ML methods for automated segmentation are support vector machines and tree ensembles algorithms, which have shown promising results for thoracic, abdominal, and pelvic tumors and normal tissue segmentation [5]. CNNs of U-net architecture are commonly used in DL for segmentation tasks. The U-Net is a CNN that was created by Ronneberger et al. [16] for biomedical image segmentation and has proved to be successful. The architecture is build upon the fully convolutional network [17] and was modified and extended to work with fewer training images and to yield more precise segmentations.

2.3 Artificial intelligence

The idea of AI came into existence in the 1950s, and the term was first coined in 1956 [18]. That said, the concept of AI is not very new, although it did not gain much popularity until recently. The reason for this is that large amounts of data did not exist earlier, and the data that existed was not good enough to predict accurate results. However, in the contemporary era of big data, there is a significant increase in data volumes and advanced algorithms, and together with improvements in computer power and storage, this is making AI one of the fastest-growing areas of technology today.

AI allows computers to simulate human intelligence by reproducing human behavior and nature learned from the surrounding environment. AI aims to give computers the ability to learn and potentially improve the performance of their tasks. The term AI is defined in many ways. However, a commonly used definition was provided by Elaine Rich in 1983, describing AI as "the study of how to make computers do things at which, at the moment, people are better" [19]. Included in AI is both ML and DL. DL is a is a more advanced type of ML. Figure 5 shows in a simple way how these three concepts are related to each other.

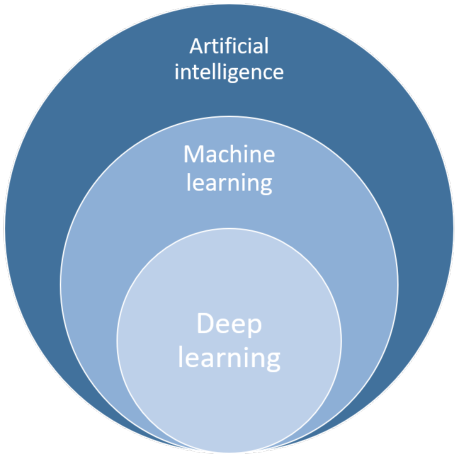


Figure 5: Diagram showing how AI, ML, and DL relate to each other. AI is a technique that enables computers to mimic human behavior. ML is a subset of AI that enables computers to learn without being explicitly programmed to do so. DL is a subset of ML again, which uses DNNs to learn many levels of abstraction, allowing the computer to train itself.

2.3.1 Machine learning

ML allows programs to learn and make decisions based on their past data. Arthur Samuel is one of the pioneers of ML, and in 1959, he described ML as the study of algorithms and statistical models that machines use to perform tasks without having to be explicitly programmed for it [20, 21]. In other words, an ML system is trained rather than explicitly programmed. Such a system can deal with large complex datasets, and when presented with multiple examples relevant to a task, it can find statistical structure in these examples that eventually allows the system to come up with rules for automating the task [22].

Types of machine learning

ML algorithms are commonly subdivided into supervised and unsupervised learning [20]. Figure 6 shows the two common types of ML and examples of the techniques. The main difference between the two types is that in supervised learning, the model is trained using labeled data, meaning that the data is already tagged with the correct answer. Unsupervised learning, on the other hand, deal with mainly unlabeled data.

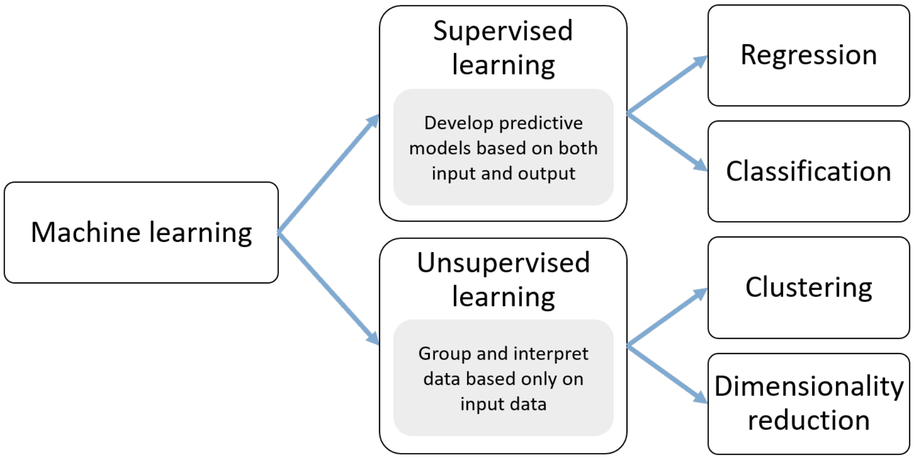
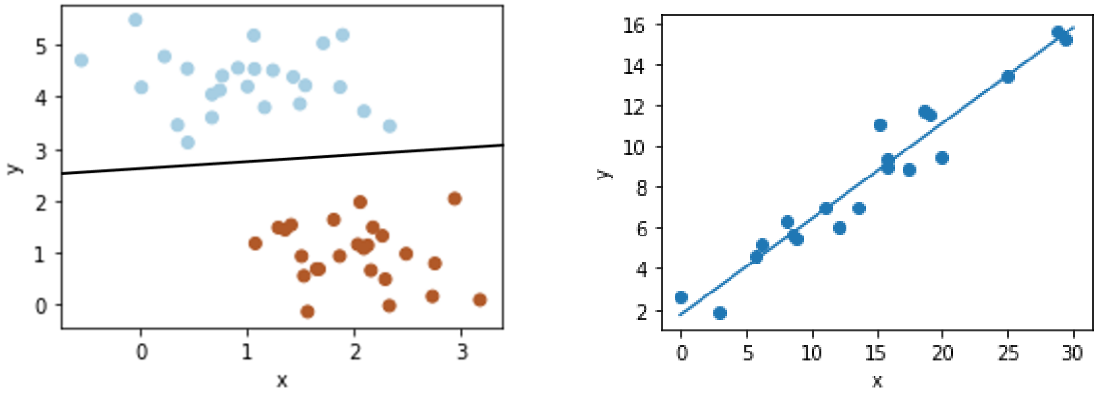


Figure 6: ML is broadly divided into two main categories: supervised and unsupervised ML. Regression and classification are two types of supervised ML techniques, and clustering and dimensionality reduction are two types of unsupervised learning techniques.

The main goal of supervised learning is to train a model from labeled data in order to make predictions about unseen or future data. With a known input and a known output, the goal is to learn a mapping from the input to the output. An example of supervised learning is classification, which is typically applied in medical imaging and image recognition. Regression is another common supervised learning technique. In classification problems, the variables are categorized to form the output, while in regression problems, the output variables are set as real numbers [20].

Figure 7 shows two ML tasks. The diagram to the left shows a collection of two-dimensional data, colored according to two different class labels. In this task, a classification algorithm can be used to draw a dividing boundary between the two clusters of points, as shown in the figure. By drawing this separating line, the model learns to make generalizations about new data: The algorithm can now predict whether a new, unlabeled point is a blue or orange point [23]. The diagram to the right shows a regression task: a simple best-fit line to a set of data. This is also an example of fitting a model to data, and by learning from the training data, the model can be used to predict the y-value when given an x-value.

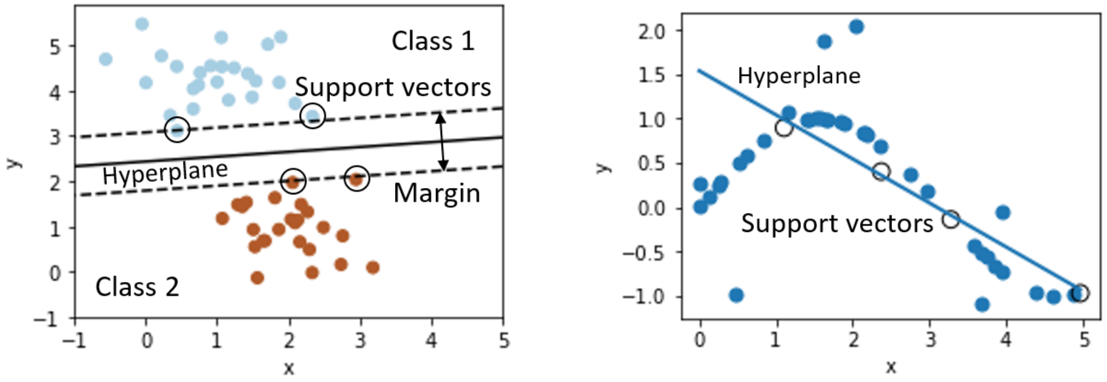


(a) A classification problem

(b) A regression problem

Figure 7: Examples of two simple supervised ML tasks.

Support vector machines are supervised learning models used for classification and regression [23]. The idea behind the support vector machines is simple: The algorithm tries to make a decision boundary in such a way that the separation between the two classes is as wide as possible. First, it finds the points closest to the line from both the classes. These points are the support vectors. Next, the distance between the line and the support vectors is computed. This distance is called the margin, and the goal is to maximize it. The hyperplane for which the margin is maximum is the optimal hyperplane. The only points that will affect the location of the hyperplane is the points either laying on the margin or violating it. Support vector classification (SVC) is a method that is based on the creation of such a hyperplane. Figure 8a shows a plot of the support vectors in linear SVC. The method of SVC can be extended to solve regression problems [23]. This method is called support vector regression (SVR), and an example of linear SVR is shown in Figure 8b.



(a) Linear SVC

(b) Linear SVR

Figure 8: The principle of the linear SVC and SVR methods. The hyperplane is optimized to separate the data into two classes in SVC and to find the line that best approximates all the individual data points in SVR.

Unsupervised learning is dealing with unlabeled data or data of unknown structure. Here, only input samples are given to the learning system, and data is grouped and interpret based solely on this input data. In this case, the goal is not to predict a variable; instead, regularities and patterns in the input data are investigated. Clustering and dimensionality reduction are examples of this type of prediction. Clustering predictions are made by finding clusters or grouping of the input, while dimensionality reduction refers to methods that reduce data from a higher dimensional space to lower dimension by using the principal components [20]. Figure 9 shows an example of a clustering problem. The algorithm aims to automatic group similar objects into sets, such that the data points in the same group are more similar to each other than to those from different groups [23].

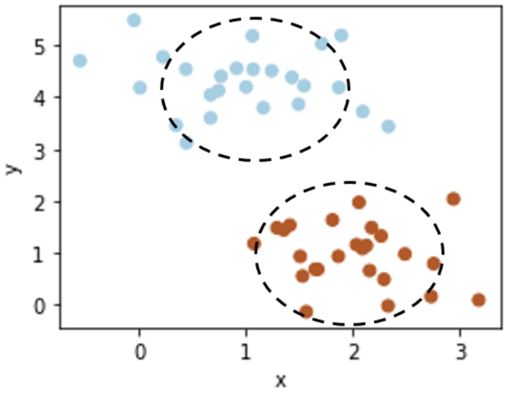


Figure 9: Example of a simple clustering problem, which is an unsupervised ML task.

Building a machine learning system

The process of optimizing an algorithm is called training. It is in this process that the model learns relevant patterns of the input samples. Figure 10 shows a diagram illustrating a typical workflow for using ML in predictive modeling.

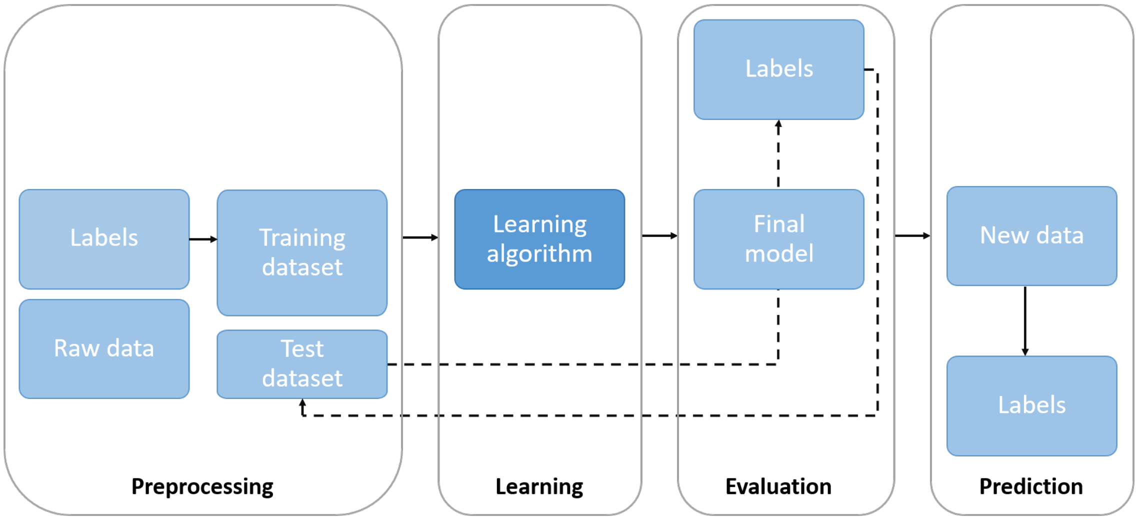


Figure 10: Typical workflow for using ML in predictive modeling [24]. Before training and selecting a predictive model, pre-processing is necessary to get data into shape. When satisfied with its performance, the model can be used for predicting new, unseen data instances.

Before training, the raw data needs to be pre-processed to get into the form and shape that is necessary for achieving the optimal performance of the learning algorithm. Further, the dataset is divided into separate training and test sets. The training set is used to train and optimize the model, while the test set is used as a final evaluation of the model and contains unseen samples. When satisfied with the model's performance, the model can be used to predict new, future data [24].

In addition, one can divide the training set further into training and validation subsets to validate the proposed weights after the training and observe how the model performs on new, unseen data before the final evaluation. Following this, one can decide whether further training of the algorithm is necessary or not, depending on how well the performance is on the validation set [24].

2.3.2 Deep learning

DL is a ML technique where algorithms train themselves and perform tasks by using deep neural networks (DNNs). A DNN is type of artificial neural network (ANN). ANNs are sets of algorithms designed to interpret sensory data and recognize patterns, inspired by the functionality of the human brain cells. But, unlike a biological brain where any neuron can connect to any other neuron within a certain distance, the ANNs have discrete layers, connections, and directions of data propagation. Like in ML, the learning can be categorized as supervised, semi-supervised, and unsupervised [25].

DNNs can consist of numerous layers of neurons that each evaluate its input signals and supply a proceed signal to the next layer. The neurons are mathematical functions. Each neuron assigns a weighting to its input, describing the importance of the connection relative to the other connections. Prior to training, it is common to set all weights to zero or small random numbers. Then, when training the DL network, one iterates through the network several times, and for each training sample, the output is computed, and the weights of the connections are updated. The networks require many training samples until the weightings of the neuron inputs are tuned precisely [24].

The neurons are typically organized into multiple layers, especially in DL. The layer that receives external data is the input layer, and the layer that produces the result is the output layer. Between them are zero or more hidden layers. Between two layers, multiple connection patterns are possible. When all neurons in a layer are fully connected with all neurons in another layer, the layer is fully connected. Figure 11 is an example of a network consisting of fully connected layers and illustrates how a DNN can look like. The network has one input layer, two hidden layers, and one output layer. The units in the first hidden layer are fully connected to the input layer, and the output layer is fully connected to the second hidden layer. Since this ANN has more than one hidden layer, it is called a DNN. Layers can also be pooling, where a group of neurons in one layer connect to a single neuron in the next layer, thereby reducing the number of neurons in that layer [26].

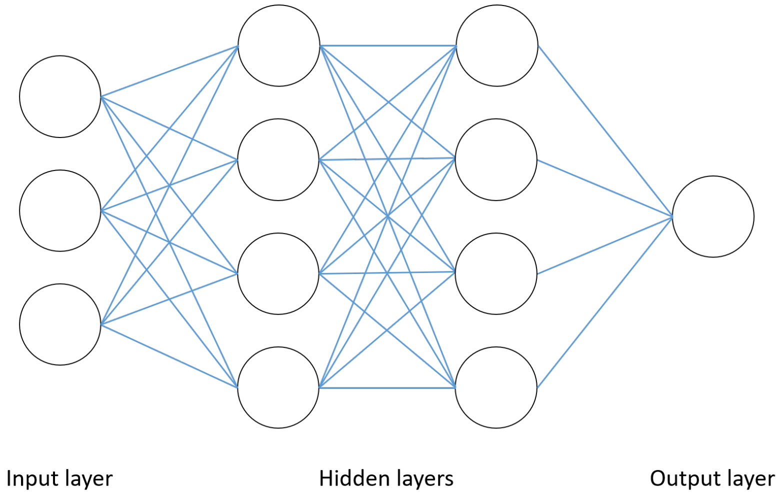


Figure 11: An example of an DNN with two hidden layers. The circles represent activation units, and the number of activation units in the first layer depends on the number of variables in the input data. The final layer is the output signal from the network. In between are the hidden layers, where the information is processed. The blue lines represent connections, each with a given weight.

Figure 12 illustrates how the architecture of a DNN can be. The input samples and the corresponding weights are combined to compute the net input. The net input is then passed on to the activation function, which, based on the information from the network, computes a prediction for the given sample. During the learning phase, this output is used to calculate the error of the prediction and update the weights [24].

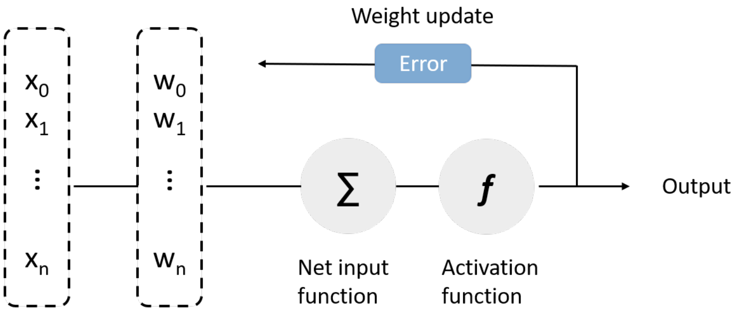


Figure 12: Diagram illustrating how the architecture of a DNN can be. The inputs of sample x and the corresponding weights w are processed through a net input function and an activation function before the model obtains an output, and the weights are updated.

Activation functions are mathematical equations that determine whether a neuron should be activated or not, based on whether its input is relevant for the model's prediction. Many different activation functions exist. It can be a simple step function that turns the neuron output on and off, depending on a rule or threshold. Or it can be a transformation that maps the input signals into output signals that are needed for the neural network to function. For instance, a linear activation function takes the form

$$f(z) = \mathbf{w}^T\mathbf{x} = a, \tag{2.1}$$

where z is the net input computed with the transposed weights vector $\mathbf{w}^T$ and the samples vector $\mathbf{x}$ [24]. The scalar a is the resulting activation, which is forward propagated to the next layer. This type of function takes the inputs, multiplied by the weights for each neuron, and creates an output signal proportional to the input. Another example of an activation function is rectified linear unit (ReLU), which is defined as:

$$f(z) = max(0, z). \tag{2.2}$$

ReLU sends an activation signal to the next neuron layer only if the input value is above zero, as shown in Figure 13. It introduces non-linearity and allows the network to converge very quickly, making it computationally efficient [24].

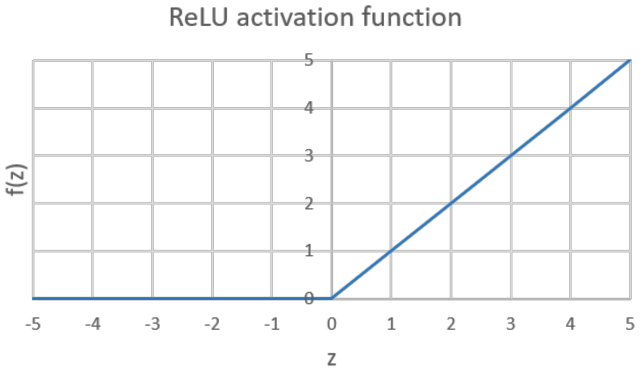


Figure 13: ReLu activation function where z is the net input, and f(x) is the activation function.

Convolutional neural networks

A CNN is a class of DNNs, most commonly applied to analyzing visual images. CNNs are neural networks that use convolution instead of general matrix multiplication in at least one of their layers [27]. A CNN consists of an input and an output layer, as well as multiple hidden layers. Typically, CNNs are composed of several convolutional layers and pooling layers that are followed by one or more fully connected layers at the end [24]. The activation function is commonly a ReLU layer.

A key to performance for any ML or DL algorithm is to successfully extract relevant features. Neural networks can automatically learn the features from raw data that are most useful for a particular task. The early layers, the ones right after the input layer, extract low-level features. Deep CNNs combine these low-level features in a layer-wise fashion to form high-level features. For example, when dealing with medical images, low-level features, such as lines and edges, are extracted from the earlier layers, which are combined together to form high-level features, such as object shapes like target volumes or OARs [24]. Layering of convolutions allow the network to account for increasingly more complex patterns.

2.3.3 Artificial intelligence in radiation oncology

AI is rapidly transforming many areas of technology. In the field of radiation oncology, efforts have been made to advance the possibilities of using AI systems to facilitate and improve the efficiency of the radiotherapy workflow process, which was illustrated in Figure 1. AI, with the use of ML and DL, have been applied in almost every part of this process. In particular, AI has been proposed for automatic organ segmentation and automatic plan generation [8].

For organ segmentation, several commercial auto-segmentation algorithms already exist. However, the underlying technology often relies on an atlas-based and model-based strategy rather than utilizing AI. The performance of atlas-based methods depend highly on the type of structure, showing better results for high-contrast organs while struggling with soft tissue organs [28]. Further, the use of model-based segmentation is generally limited to specific organs. Currently, such auto-segmentation tools in treatment planning are most commonly viewed as an efficient tool for the clinicians to provide them with a good starting point for review and adjustment [2].

However, recent advances in DL have come up with faster and more accurate solutions for auto-segmentation. For example, Lustberg et al. [29] compared the aspects of contouring ROIs manually with atlas-based and DL-based contouring for lung cancer patients, showing promising results for DL. The DL contouring outperformed the atlas-based contouring for several structures and in time saved. Men et al. [30] proposed a DL method using CNNs for auto-segmentation of ROIs in rectal cancer. The results showed that this method could improve the consistency of contouring and increase the efficiency of the radiotherapy workflow. Tong et al. [31] developed a DL method using CNNs for segmentation of OARs in head and neck cancer radiotherapy. This method showed competitive performance, and it took shorter time to segment multiple organs in comparison to state-of-the-art method.

While it is clear that each of the methods described above are useful, all remain within the domain of research and have not been made available commercially. However, vendors of modern treatment planning systems have recently integrated AI in their software. For example, RayStation 8B (RaySearch Laboratories AB, Stockholm, Sweden) was the first treatment planning system to incorporate ML applications. This system uses a classical ML method based on random forest for automatic plan generation and DNNs for organ segmentation. Also, the first-ever patient treatments generated using ML in RayStation took place in May 2019. Another commercial software utilizing AI in its applications is Eclipse$^{\text{TM}}$ v16 (Varian Medical Systems, Palo Alto, California). This system uses an atlas-based ML model in which a group of representative plans is used as a base model. This system also include the first clinical application of ML in proton treatment planning. However, the commercially available products utilizing AI are not frequently used in clinical practice.

Even though DL solutions shows promising results compared to existing solutions for auto-segmentation, most remain within the domain of research. However, with continuous ongoing research, it is reason to believe that AI-based methods will have a significant role in generating segmentations in near future, at a much faster and more consistent manner than what is possible to do at present. Further, it is reasonable to expect increased availability of commercial AI-based auto-segmentation tools for radiotherapy treatment planning over the next years; and with this, an increased acceptance and implementation of AI-based auto-segmentation tools in clinical practice.

2.4 Breast cancer

Breast cancer is the most common type of cancer in women worldwide, with 2,1 million new cases diagnosed in 2018 [32]. In Norway, breast cancer contributes to about 22 % of all cancer cases affecting women, and more than 3500 patients are diagnosed with this disease every year [33].

2.4.1 Anatomy and physiology

Breast cancer develops from the tissues of the breast. Figure 14 gives an illustration of the female breast. Each breast has 15 to 20 sections called lobes, and each lobe comprises many smaller sections called lobules, at the end of which are glands that produce milk in response to hormones. The lobes, lobules, and glands are linked by thin tubes called ducts. The most common type of breast cancer is called ductal carcinoma and begins in the cells of the ducts. Another type of breast cancer is lobular carcinoma, which begins in the lobes or lobules. Invasive breast cancer is breast cancer that has spread from where it began in the breast ducts or lobules to surrounding normal tissue. Breast cancer occurs in both men and women, although male breast cancer is rare [34].

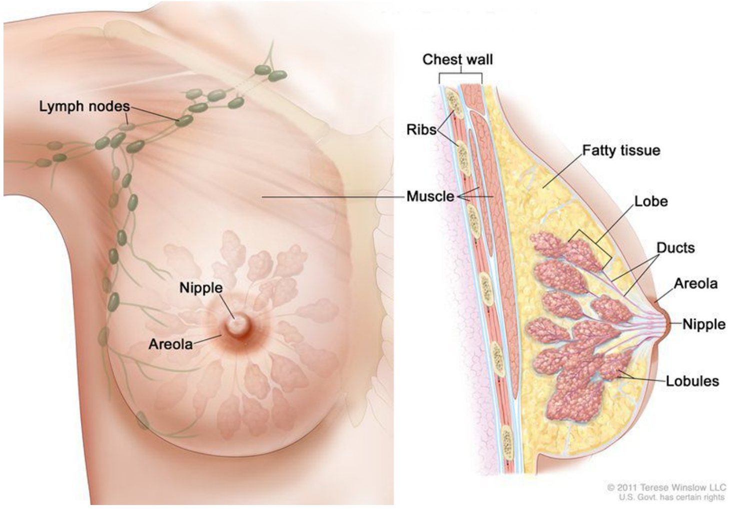


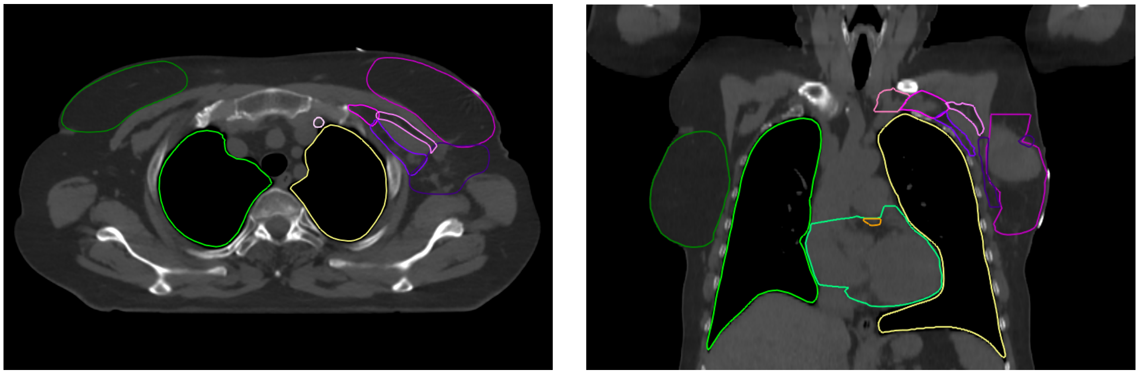
Figure 14: Anatomy of the female breast—courtesy of [34]. The nipple and areola are shown on the outside of the breast. The lymph nodes, lobes, lobules, ducts, and other parts of the inside of the breast are also shown.

2.4.2 Treatment modalities

Breast cancer is treated in different ways, depending on the size of the tumor, the characteristics of the cancer cells, and whether the cancer cells have spread to nearby lymph nodes. Alongside surgery, chemotherapy, and hormone treatment, radiotherapy is commonly used for breast cancer treatment. Most breast cancer tumors can be removed with surgery. In the majority of the cases, breast-conserving surgery is performed, where only the tumor with nearby tissue is removed. If the tumor is large compared to the breast, or there are multiple tumors spread around the mammary gland, the entire breast is removed. Radiotherapy is given after breast-conserving surgery to remove possible remaining cancer cells. When the entire breast is removed, radiotherapy is given in the case of spread to lymph nodes, or if any cancerous tissue was missed during surgery. If the disease cannot be cured, radiotherapy can limit the disease and provide palliation. In addition, chemotherapy is used to prevent spread and to reduce the risk of cancer recurrence. Further, some types of breast cancers are affected by hormones, and hormone therapy is then used mainly to prevent recurrence [35].

Radiotherapy

Most breast cancer patients receive radiotherapy treatment following surgery. The relevant target volumes are the breast, the chest wall, and the regional lymph nodes, as shown in Figure 15. For patients with locoregionally advanced disease, the following regional lymph nodes are considered in addition to the breast: the axillary nodes, the supraclavicular region, the interpectoral nodes, and the internal mammary nodes region [36]. Recommended radiation doses are hypofractionated regimes consisting of 40 Gy in 15 fractions or conventionally fractionated regimes consisting of 50 Gy in 25 fractions. Additionally, a boost to the tumor bed is given to patients younger than 50 years old after breast-conserving surgery [37].



(a) Transverse plane

(b) Coronal plane

Figure 15: Example of ROIs relevant for breast cancer radiotherapy. The target volumes include the breast and different regional lymph nodes and are shown in purple colors. The OARs are shown in green and yellow colors and include the heart, the lungs, the contralateral breast, and the LAD.

Relevant OARs to consider for breast cancer patients are also shown in Figure 15 and include the heart, the lungs, the left anterior descending coronary artery (LAD), and the contralateral breast. If regional lymph nodes are included in the target volume, medulla spinalis and plexus brachialis should also be considered [37]. These structures are routinely defined and contoured on the patient scan by a physician or radiation therapist.

3 Materials and method

3.1 Evaluation of a DL thorax model

A DL-based model for auto-segmentation of organs in the thorax region, implemented in a commercial treatment planning system, was evaluated by generating segmentations for the heart, the left and right lungs, the spinal cord, and the esophagus.

3.1.1 The model

The thorax model in RayStation 9A (RaySearch Laboratories AB, Stockholm, Sweden) is an organ segmentation model based on DL. The model is based on lung cancer patients, and it is suitable for CT image modality and patient position head first-supine. The model came pre-trained in RayStation and was trained using supervised learning on annotated images, starting from a randomly initialized model. The model was trained with 65 segmented image sets, originating from Centre Oscar Lambret (Lille, France). The training data was augmented by rotations, translations, and elastic deformations.

The model algorithm is a CNN of U-net architecture, and the originator of the scripting environment is RaySearch. The DL segmentation algorithm is a voxel classifier using DNN architectures with multiple hidden layers to learn features from a training set by modeling complex non-linear relationships. Each voxel in the image is classified as belonging either to unspecified tissue or to a specific structure. The algorithm is trained on a large number of segmented images to learn how to classify the voxels. The specific network used is a three-dimensional CNN of U-net architecture, which can combine image features on different levels of abstraction to generate a segmentation map. Figure 16 shows a simplified illustration of the network.

This type of network combines encoding and decoding paths with skip-connections to concatenate features from the encoding to the decoding layers, allowing the network to work with features at different resolutions. The number of features available to the algorithm is predefined, but the features themselves are not. Instead, the algorithm learns the most important features from the dataset used during the training of the model. With a constant number of features, the algorithm can learn from an unlimited number of training cases without affecting the size or runtime of the model. In addition, the algorithm is graphics processor unit (GPU)-powered, which allows for fast segmentation [15].

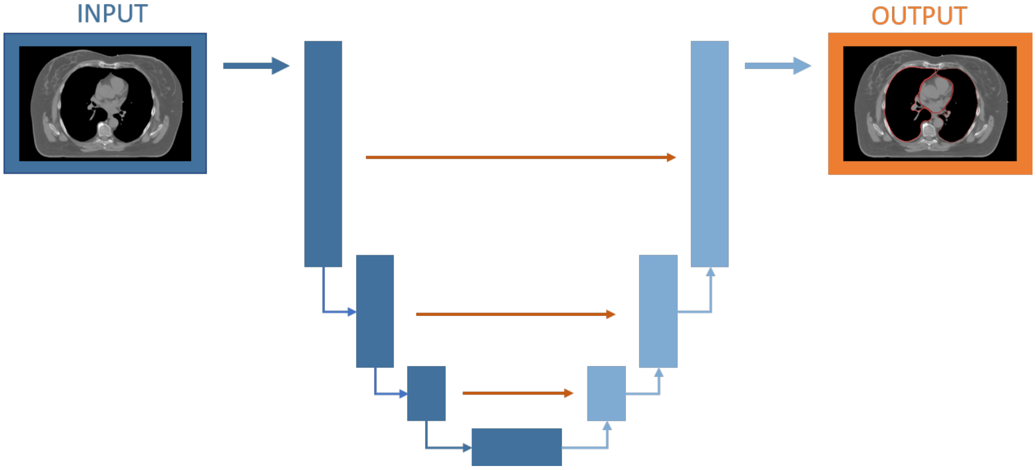


Figure 16: A simplified illustration of the CNN of U-net architecture. Each block represents a set of consecutive convolutional layers, and the orange arrows are skip connections. The output of the blocks is downsampled on the left side of the network and upsampled on the right side.

3.1.2 Patient data

All use of patient data in this study was applied for and pre-approved by the Regional Committees for Medical and Health Research Ethics (REK Midt ref. 92685). All patients were diagnosed with left-sided breast cancer and previously treated with external photon beam radiotherapy at St. Olavs Hospital, using deep inspiration breath hold. This is a controlled breathing technique in which the patient performs a breath hold during treatment. Radiotherapy planning and treatment were performed according to the protocol at St. Olavs Hospital, which includes several patient fixation steps. Breast boards, most commonly WingSTEP from ELEKTA, were used to enable easy positioning, precise repositioning, and patient comfort during treatment. When needed, a 10 degrees elevation cushion was put under the breast board to lift the upper body and thereby facilitate breathing. The arms were positioned above the head and out of the treatment fields, and a head rest, knee pillow, and arm support could be used to support the neck and stabilize the back and pelvis.

Radiotherapy planning CTs for 20 patients treated in 2019 were selected for testing the DL thorax model in RayStation 9A. All patients received locoregional treatment except 2, which received only breast irradiation. A hybrid technique that incorporates both conventional fields and VMAT was used in 17 of the cases, while in the last 3 cases, the patients were treated with full VMAT. Further, 8 of the patients were treated with 2,67 Gy x 15, and 12 of the patients were treated with 2 Gy x 23. All segmentations were previously clinically approved and used in the delivered radiotherapy plans.

3.1.3 Evaluation

For each patient, the CT images with segmented ROIs, originally planned in RayStation 6, were anonymized and exported to a non-clinical installation of RayStation 9A. The original planning CT was used to retrospectively create new segmentations of the heart, the left and right lungs, the spinal cord, and the esophagus for each patient using the DL thorax model in RayStation. RayStation supports scripting, and scripts were written in Python to extract data for quantitative analysis. A script for extracting dose values is attached in appendix A, and a script for computing quantitative measures for comparison of ROIs is attached in appendix B. The overall segmentation time for the AI structures were measured for each patient. The DL contouring used a graphics card to perform the calculations, and the GPU used was a NVIDIA Quadro K5200 with 8 GB of GDDR5 memory.

Dosimetric analysis

When evaluating treatment plans in radiotherapy, several parameters are used to determine whether a treatment plan gives good enough dose coverage to the tumor and good enough sparing of the OARs. DVH parameters are commonly used to evaluate treatment plans, together with inspection of the three-dimensional dose distribution. It is therefore interesting to see whether the differences in manual and automatic contouring affects the calculated OAR doses.

The segmentations obtained with the DL thorax model were compared to the manual segmentations in terms of dose to the heart and lungs. For the heart, the average dose was considered, and for the lungs, the average dose and the volume that receives either 18 Gy or 20 Gy, depending on the used fractionation regime, were considered.

The dose evaluation criteria for the lungs are dependent on whether the patient has received locoregional radiotherapy or not. The criteria considered in this study are based on the clinical goals used at St. Olavs Hospital and are summarized in Table 1. For irradiation of the breast and regional lymph node ares, the following criteria apply:

- For 2 Gy x 25 fractions, less than 35 % of the lung should receive 20 Gy (V20 $\leq$ 35 %).
- For 2,67 Gy x 15 fractions, less than 35 % of the lung should receive 18 Gy (V18 $\leq$ 35 %).

For irradiation of the breast only, the criteria are:

- For 2 Gy x 25 fractions, less than 15 % of the lung should receive 20 Gy (V20 $\leq$ 15 %).
- For 2,67 Gy x 15 fractions, less than 15 % of the lung should receive 18 Gy (V18 $\leq$ 15 %).

Table 1: Dose evaluation criteria for lungs.

	Locoregional	Breast only
2 Gy × 25 fractions	V20 < 35 %	V20 < 15 %
2,67 Gy × 15 fractions	V18 < 35 %	V18 < 15 %

3.2 Training and testing of ML models

ML-based models for auto-segmentation were trained and tested for contouring of the sternum, the left breast, and the heart. The sternum was chosen because this is a structure routinely contoured at St. Olavs Hospital to help with matching of the setup images before treatment. In addition, the sternum is a well-defined structure and was therefore assumed to be suitable for ML-based contouring and relatively few training dataset would be required.

3.2.1 The algorithm

The ML method is developed at the Department of Physics, NTNU (Trondheim, Norway) for automatic detection of image structures, originally for the use of MR images. From before, the model has been trained for automatic segmentation of the tumor volume for rectal cancers [38]. It uses linear SVC to do a voxelwise classification on the images to separate the structure and the normal tissue. In general, each voxel in the image is classified as belonging either to unspecified tissue or to a specific structure. The model is developed in Python version 3.7.5, and the main libraries used are NumPy, SimpleITK, Scikit-learn, and Dask. NumPy is the core library for scientific programming in Python and is used for creating multi-dimensional array objects. SimpleITK is a simplified version of the Insight Segmentation and Registration Toolkit (ITK), which provides a broad set of tools required for image analysis. Scikit-learn is tool for predictive data analysis and includes a large collection of ML algorithms. Dask is a library for parallel computing, making it possible to work with large datasets that exceed the memory of the computer.

The model starts by splitting the dataset into training and test sets. The training set is put into the ML algorithm together with the corresponding class labels, which generates a model that takes the test set as input and outputs predicted labels. The model is then evaluated by comparing the predicted labels to the real labels of the test set. The ML algorithm used in the models is the `sklearn.linear_model.SDGClassifier` from the Scikit-learn library together with the wrapper function `dask_ml.wrappers.Incremental` from the Dask library.

3.2.2 Patient data

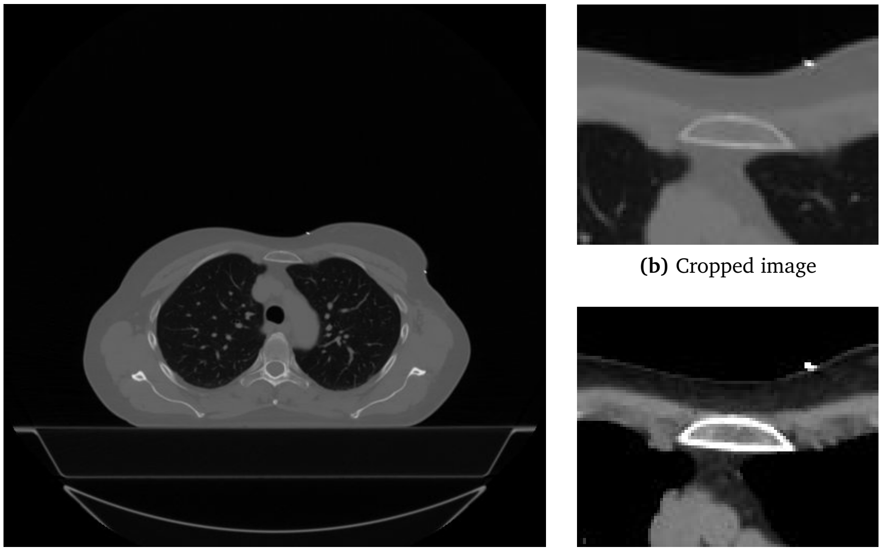
Radiotherapy images from 30 patients were used for training and testing the ML models. These are from patients participating in the COBRA study [39], which include left-sided breast cancer patients. All patients were treated with external photon beam radiotherapy at St. Olavs Hospital between 2017 and 2018, using deep inspiration breath hold. Radiotherapy planning and treatment were performed according to the protocol at St. Olavs Hospital, as described in section 3.1.2. The patients are CT scanned with the same protocol, but field-of-view vary to some extent, depending on the size of the patient. The standard is 512 pixels both in x and y directions, but due to varying field-of-view, the exact pixel size is different for different patients. In the longitudinal direction, the scan is taken between the angle of the mandible and the bottom of the lungs, meaning that the number of slices also vary. The slice thickness and distance is fixed at 3 mm.

3.2.3 Training and testing

For each patient, planning CTs, together with the segmented target volumes and OARs, were anonymized in the clinical version of RayStation 8B and exported to a non-clinical installation of RayStation 8B. The sternum was manually segmented and exported together with the clinical segmentations of the left breast, the heart, and the CT images as DICOM files. The image data was then converted to NIfTI format, as this is the format the ML algorithm is built to work with. The NIfTI format is a common format used to store MR imaging data, and it is made up of a header file containing the metadata and a data file containing the image data. This data was then used to train the ML models to automatically segment the different ROIs. The models were trained with both 20 image series and 30 image series to see if increasing the amount of training data could improve performance of the models. Scripts were written in Python to calculate data for quantitative analysis, and functions from the SITK library were mainly used for this.

Pre-processing

Before training the models, the images were processed. The different images were modified in a similar manner as shown in Figure 17. The images were cropped to a fixed region around the structure of interest to reduce the size of the data and obtain a more balanced dataset. This was done such that there was a 15 mm margin outside the largest extent of the ROI amongst all the slices. This could simulate the process of a physician that draws a box around the ROI to assist the classification. New, modified images were also created by changing window/level and added to the training dataset to see if they alone or in combination could improve the results.



(b) Cropped image

(a) Original image

(c) Cropped and modified image

Figure 17: Example showing how the images were processed before they were used as input data to the ML model for the sternum. The image before (a) and after cropping (b) and changing window/level (c) is shown.

Cross-validation

In order to train and evaluate the models, the data needed to be divided into training and test sets. Leave-one-out cross-validations were used to perform this task. The principle of this method is illustrated in Figure 18. With this method, each patient is used as test set in turn, while the rest of the patients are used for training. The number of iterations will then equal the number of patients, making this method suitable for a small dataset.

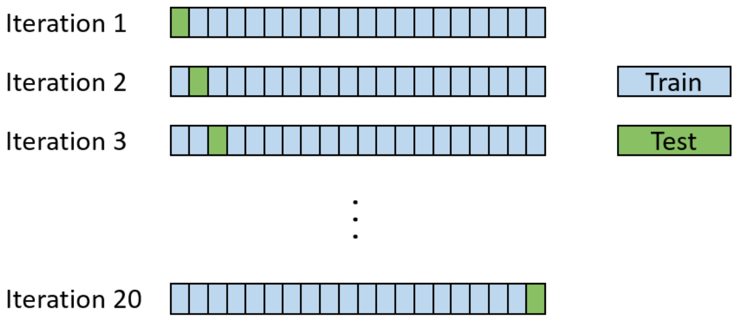


Figure 18: Leave-one-out cross-validation on a group of 20 patients. Each square represent a patient, and each patient is used for testing (🟩) in one iteration and for training (🟦) in the remaining iterations.

Post-processing

To improve the results, the images were processed after training the model as well. Morphological operations were applied to the predicted masks to remove noise and fill holes. Different approaches were investigated for the different ROIs. One approach was to remove areas smaller than a chosen number of voxels for each slice. For the sternum this limit was 100 voxels, while areas smaller than around 1000 and 2000 voxels were removed for the breast and heart, respectively. Another approach was to look at all the image slices for each patient put together and remove volumes smaller than a chosen number of voxels. Other approaches included binary morphological opening and closing of the images to remove small structures or fill small holes.

3.3 Description of methods used for comparison

The evaluation of the DL thorax model was based on segmentations of the heart, the lungs, the esophagus, and the spinal cord. The AI segmentations of the heart and the lungs were compared to the clinical segmentations, and a quantitative and clinical evaluation were performed. The esophagus and the spinal cord were not contoured manually, and these AI structures were therefore evaluated with only a clinical evaluation. For the ML models, the evaluation was based on segmentations of the sternum, the left breast, and the heart, and a quantitative evaluation was performed.

3.3.1 Quantitative evaluation

Evaluation of segmentation results is most commonly performed with overlap methods, which estimate the overlap of two volumes as a fraction of their total volume. The most common overlap method is the Dice similarity coefficient (DSC). Another standard measure is the Hausdorff distance (HD). Both methods are useful measures for the geometric quantification of segmentation similarities [40] and were therefore used to analyze the segmentations obtained with the different auto-segmentation methods. For evaluation of the DL thorax model, the DSCs, the 75-, 90-, 95- and 100-percentile HDs, and the average HDs (AVDs) for the AI and clinical segmentations were calculated. The 75-, 90-, 95- and 100-percentile HDs are denoted H75, H90, H95 and H100, respectively. For evaluation of the ML models, the DSCs and the different HD values for the AI and manual segmentations were calculated.

Dice similarity coefficient

The DSC is a simple spatial overlap index and reproducibility validation metric, first proposed by Dice in 1945 [41]. It is the metric most frequently used in literature to quantify the spatial overlap between two binary segmentation results [2]. Given two volumes of interest, X and Y, the DSC is defined as:

$$\text{DSC} = \frac{2|X \cap Y|}{|X| + |Y|} \tag{3.1}$$

where X and Y are the two volumes under comparison, and $X \cap Y$ is the union of the two volumes, as illustrated in Figure 19. Using the definition of true positive (TP), false positive (TP), and false negative (TP), this can be rewritten as:

$$DSC = \frac{2TP}{2TP + FP + FN}. \tag{3.2}$$

The value of the DSC ranges from 0 to 1, where 0 indicates no spatial overlap, and 1 indicates complete overlap.

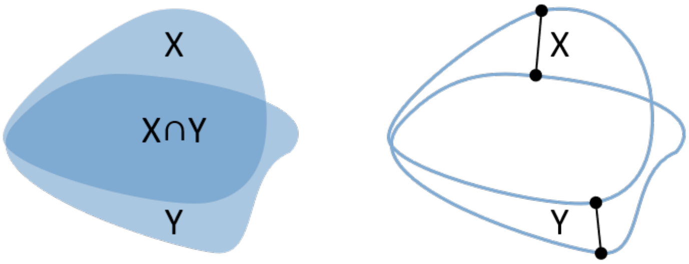


Figure 19: Illustration of evaluation measures used in this study: the parameters of the DSC to the left and the HD to the right.

Hausdorff distance

The maximum distance to agreement, or the HD, measures how far two subsets of a metric space are from each other. Mathematically, it is defined as the maximum distance of a set to the nearest point in the other set, as illustrated in Figure 19. The HD between two finite point sets X and Y is defined to be

$$\mathrm{HD}(X,Y) = \max(h(X,Y), h(Y,X)) \tag{3.3}$$

where $h(X,Y)$ is the directed HD from X to Y, given by

$$h(X,Y) = \max_{x \in X} \min_{y \in Y} \|x - y\| \,. \tag{3.4}$$

Here, $\|x - y\|$ is some norm, commonly Euclidean distance [42]. Essentially, two distance transforms are computed for measuring the HD: (1) Each point on the surface of ROI X is assigned the minimum distance to a point on the surface of ROI Y, and (2) each point on the surface of ROI Y is assigned the minimum distance to a point on the surface of ROI X. HD is then given by taking the maximum. In the case of complete overlap, HD is 0.

The HD is generally sensitive to outliers. The Hausdorff quantile method is a more robust alternative to the HD, proposed by Huttenlocher et al. [43]. In this method, the HD is defined to be the qth quantile of distances instead of the maximum, so that possible outliers are excluded. The quantile q is selected depending on the application and the nature of the measured point sets.

The average distance to agreement, or AVD, is the HD averaged over all, N, points. The AVD is known to be stable and less sensitive to outliers compared to the HD [42]. It is defined by

$$\mathrm{AVD}(X,Y) = \frac{1}{2}(d(X,Y), d(Y,X)) \tag{3.5}$$

where $d(X,Y)$ is the directed average HD from X to Y, given by

$$d(X,Y) = \frac{1}{N} \sum_{x \in X} \min_{y \in Y} \|x - y\| \,. \tag{3.6}$$

3.3.2 Clinical evaluation

The segmentations obtained with the DL thorax model were also reviewed qualitatively and evaluated subjectively by a physician at St. Olavs Hospital. For the heart and lung segmentations, the AI structures were compared to the clinical structures, and the physician pointed out which segmentation was preferred for each patient and evaluated whether each segmentation was clinically acceptable or not. For the esophagus and the spinal cord, each structure was assessed as (1) the structure is good as it is, (2) the structure needs small adjustments but serves as a good starting point, or (3) the structure does not form a useful basis for further editing, and starting over again is preferable. The reviewer could also comment on each result.

3.3.3 Statistical analysis

Different statistical methods were used to visualize and analyse the quantitative evaluation results.

Test of normality

The Shapiro-Wilk test was used to assess whether the data was likely from a normal distribution or not. The null hypothesis was rejected if the p-value was less than a chosen significance level of 5 %; in that case, the conclusion was that the data was not from a population with a normal distribution. If the p-value was greater than 5 %, the null hypothesis could not be rejected, indicating that the data was normally distributed. Q-Q plots were added for visual examination to support the conclusion. A Q-Q plot is a scatter plot created by plotting two sets of quantiles against one another. If the points form a roughly straight line, this supports the assumption of normally distributed data.

Appendix C shows the results from the normality test of some of the dosimetric parameters measured for the clinical segmentations and the AI segmentations produced with the DL thorax model, as an example of how the normality tests were conducted.

Wilcoxon signed-rank test

The Wilcoxon signed-rank test was used to examine whether there was a statistically significant difference between relevant dosimetric parameters measured for the clinical segmentations and the AI segmentations produced with the DL thorax model. This is a non-parametric test and was chosen since the sample size was small, and the test statistics could not be assumed to follow a normal distribution. The Wilcoxon signed-rank test is a paired difference test and tries out the null hypothesis that the population mean ranks of two related samples are equal. The null hypothesis was rejected for a p-value less than a chosen significance level of 5 %.

Wilcoxon rank sum test

The Wilcoxon rank sum test was used to examine whether there was a statistically significant difference between the DSCs and AVDs obtained with the ML models trained with 20 and 30 image sets. This is a non-parametric test and was chosen since the sample size was small, and the test statistics could not be assumed to follow a normal distribution. The Wilcoxon rank sum test tries out the null hypothesis that two samples are likely to derive from the same population. The null hypothesis was rejected for a p-value less than a chosen significance level of 5 %.

Student's t-test

A two-sample t-test was used to examine whether there was a statistically significant difference between DSCs and AVDs for the clinically acceptable and not clinically acceptable heart segmentations for the DL thorax model. This test assumes that the test statistic follows a normal distribution. The paired t-test tries out the null hypothesis that the means of the two samples are equal. The null hypothesis was rejected for a p-value less than a chosen significance level of 5 %.

Boxplots

Boxplots were used to visualize the distribution of data. The line inside the box represents the median of the sample, and if the median is not centered in the box, this shows skewness of the sample. The box is divided in two parts where 25 % of the scores fall below the lower quartile and 75 % percent of the scores fall below the upper quartile value. The length of the box is called the interquartile range and represents the middle 50 % of the observations, and the whiskers represent scores outside the interquartile range. Outliers are observations that deviates more than 1,5 times the interquartile range and are displayed with circles [44]. The average of the sample is displayed as a cross.

4 Results

4.1 Evaluation of a DL thorax model

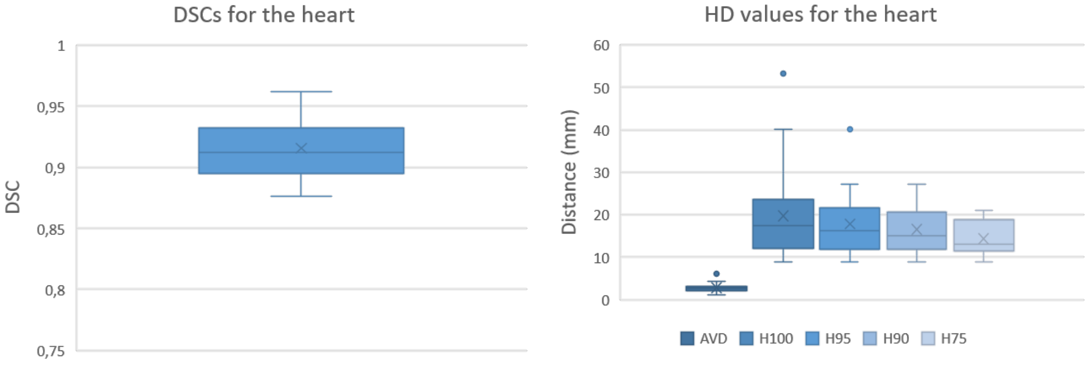
The DL thorax model used on average 3 minutes on generating AI segmentations of the heart, the lungs, the spinal cord, and the esophagus for one patient.

4.1.1 Heart

The average DSC and HD values for the heart segmentations obtained with the DL thorax model and the clinical segmentations are displayed in Table 2. Boxplots with all the DSCs and HD values for the test data, seen in Figure 20, shows that the performance of the model varied between the individual patients in the dataset for the heart segmentation.

Table 2: Average values for all metrics for the heart segmentations obtained with the DL thorax model. All HD values are in mm.

Metric	Average	STD	Min	Max
DSC	0,92	0,02	0,88	0,96
H75	14,6	4,1	8,9	21,0
H90	16,6	5,7	8,9	27,2
H95	17,9	7,9	8,9	40,2
H100	19,7	11,1	8,9	53,2
AVD	2,9	1,1	1,1	6,1



(a) DSC

(b) HD values

Figure 20: Boxplots with the DSC and different HD results for the heart segmentations obtained with the DL thorax model.

The result of the clinical evaluation is shown in Figure 21. The AI-generated segmentations passed in 42 % of the cases, in terms of clinical acceptability. The model seemed to struggle especially with the cranial part of the heart. The clinical segmentations were preferred over the AI segmentations in 89 % of the cases. However, 5 of the 19 clinical segmentations were also assessed as inadequate.

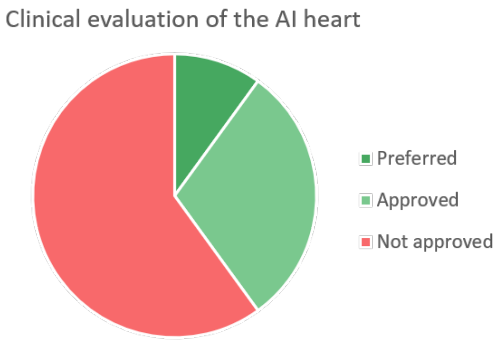
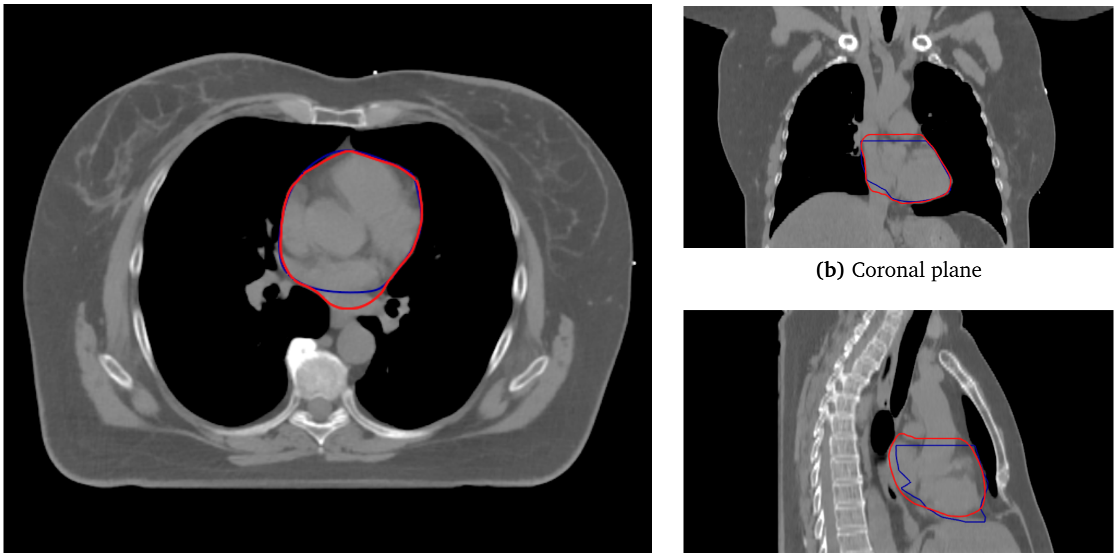


Figure 21: Clinical assessment of the heart segmentations obtained with the DL thorax model. Green indicates that the AI segmentation is clinically acceptable and red are clinically non-acceptable cases. Dark green indicates that the AI segmentation is preferred over the clinical segmentation.

Figure 22 shows an example of a clinically acceptable heart segmentation generated with the DL thorax model, together with the clinical heart segmentation. In this case, the DSC = 0,91, the H100 = 18,7 mm, and the AVD = 2,9 mm. The calculated average heart dose differed with 18,55 cGy for the AI and clinical segmentation. In this case, the AI segmentation of the heart was preferred over the clinical segmentation.

The Wilcoxon signed rank test of the paired differences (clinical - AI) in average dose to the heart calculated for the clinical segmentations and the AI-generated segmentations gave a p-value of 0,387. The median, minimum, and maximum paired differences were -4,3 cGy, -52,0 cGy, and 29,9 cGy, respectively. This means that the calculated average heart dose did not differ significantly between the AI segmentations (192 ± 81 cGy) and the clinical segmentations (192 ± 73 cGy). However, Figure 23 shows that for individual patients, the difference in segmentation resulted in a quite different heart dose; the maximum change was over 50 cGy.

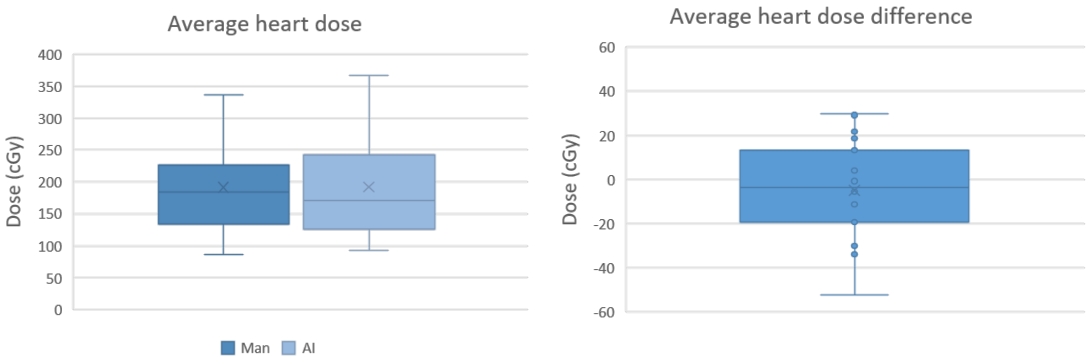


(b) Coronal plane

(a) Transvere plane

(c) Sagittal plane

Figure 22: Example of a patient with a clinically acceptable heart segmentation produced with the DL thorax model (🟥), together with the clinical heart segmentation (🟦) for one patient.



(a) Dose

(b) Dose difference

Figure 23: Boxplots with the average heart doses and average heart dose differences (clinical - AI) calculated for the clinical segmentations and the AI-generated segmentations obtained with the DL thorax model.

4.1.2 Lungs

The average DSC and HD values for the left and right lung segmentations obtained with the DL thorax model and the clinical segmentations are displayed in Table 3. Boxplots with all DSCs and HD values for the test data are shown in Figure 24 and 25, respectively. The model performed better and more uniform for lung segmentation compared to for heart segmentation.

Table 3: Average values for all metrics for the left and right lung segmentations obtained with the DL thorax model. All HD values are in mm.

	Left lung				Right lung			
Metric	Average	STD	Min	Max	Average	STD	Min	Max
DSC	0,97	0,01	0,93	0,99	0,98	0,01	0,94	0,99
H75	19,5	4,1	8,8	23,7	19,0	5,8	7,0	25,3
H90	20,5	4,4	8,8	26,7	20,5	6,3	7,0	28,5
H95	21,0	4,7	8,8	29,2	21,5	7,5	7,0	39,5
H100	21,8	5,8	8,8	36,4	22,5	8,7	7,0	42,9
AVD	1,1	0,4	0,4	2,5	0,8	0,3	0,6	2,0

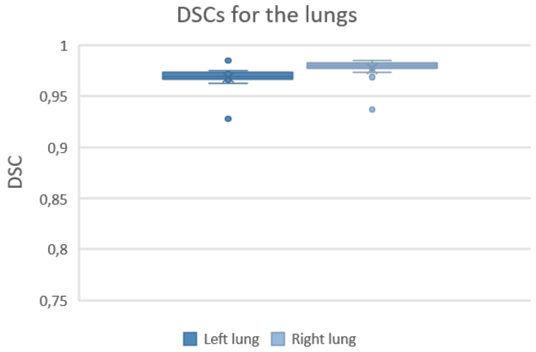
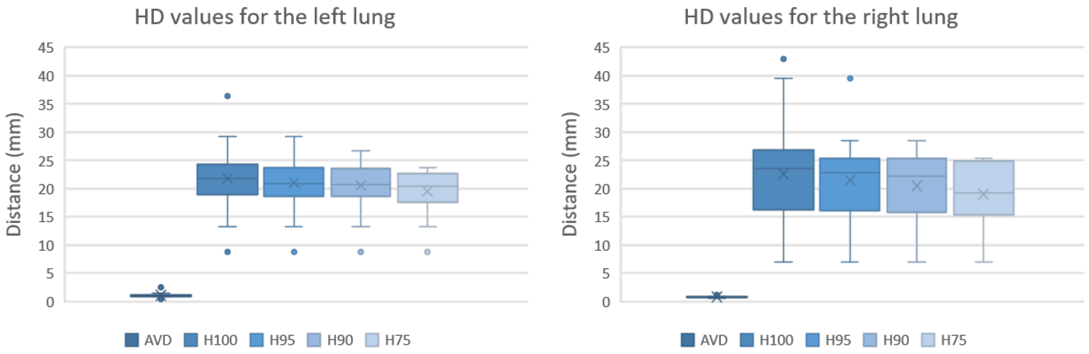


Figure 24: Boxplot with the DSCs for the left and right lung segmentations obtained with the DL thorax model.



(a) Left lung

(b) Right lung

Figure 25: Boxplots with the different HD values for the left and right lung segmentations obtained with the DL thorax model.

The result of the clinical evaluation is shown in Figure 26. Overall, the AI segmentations of the lungs were almost as good as the clinical segmentations. All of the AI-generated segmentations were assessed as clinical acceptable. However, the clinical segmentations were preferred over the AI segmentations in 70 % of the cases. In particular, the lower lung restrictions appeared to be difficult for the model.

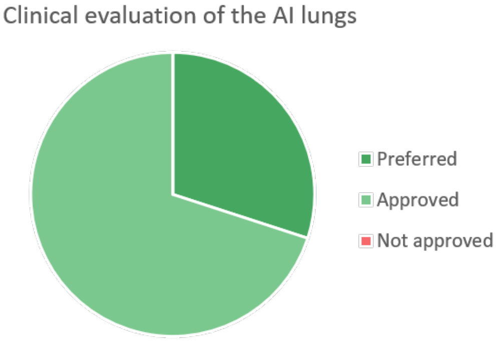
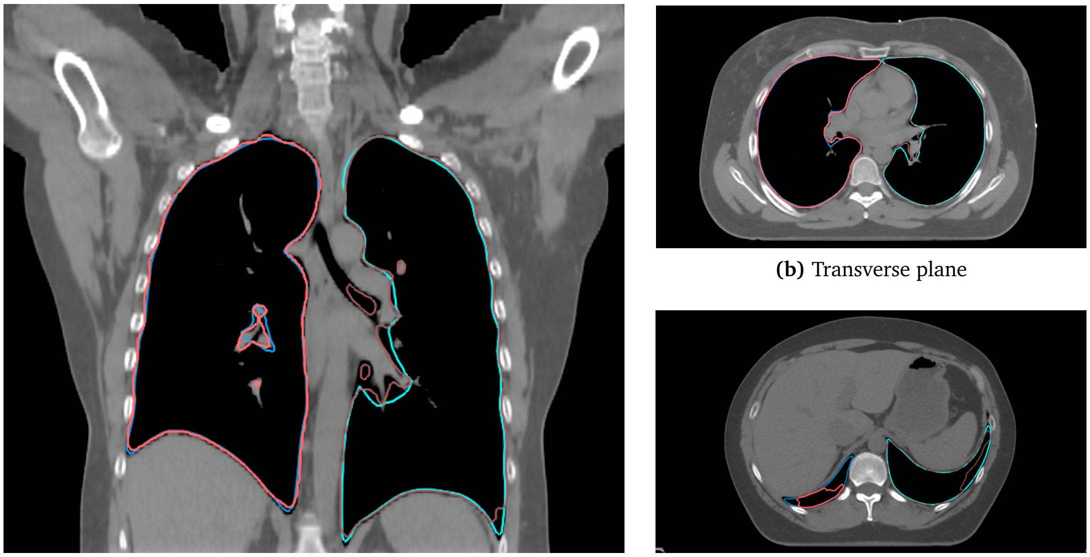


Figure 26: Clinical assessment of the lung segmentations obtained with the DL thorax model. Green indicates that the AI segmentation is clinically acceptable and red are clinically non-acceptable cases. Dark green indicates that the AI segmentation is preferred over the clinical segmentation.

Figure 27 shows a representative example of the lung segmentations generated with the DL thorax model, together with the clinical lung segmentations. For the left lung, the DSC = 0,97, the H100 = 20,3 mm, and the AVD = 1,0 mm. For the right lung, the DSC = 0,98, the H100 = 16,0 mm, and the AVD = 0,7 mm. The calculated average dose to the left and right lungs differed with 15,26 cGy and 1,60 cGy, respectively, for the AI and clinical segmentations. In this case, the AI segmentations of the lungs were deemed clinically acceptable. However, the clinical segmentations were preferred over the AI segmentations, because the lungs were worse in the basal end, as shown in Figure 27c.



(b) Transverse plane

(a) Coronal plane

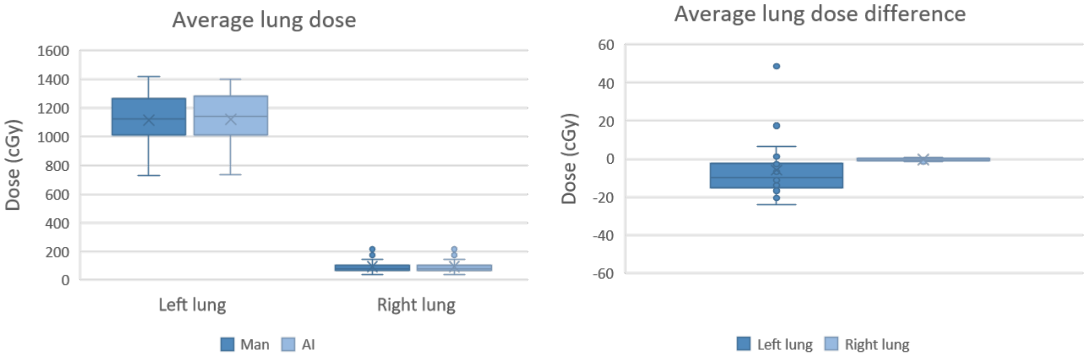
(c) Transverse plane

Figure 27: Example showing the lung segmentations produced with the DL thorax model (🟫 🟥) and the clinical lung segmentations (🟦 🟦) for one patient.

Table 4 shows the result of the Wilcoxon signed rank test of the paired differences (clinical - AI) in dose to the lungs for the clinical segmentations and the AI-generated segmentations. Average dose to the left lung differed significantly for the AI segmentations (1114 ± 196 cGy) and the clinical segmentations (1120 ± 191 cGy); as did average dose to the right lung for the AI segmentations (93 ± 42 cGy) and the clinical segmentations (94 ± 42 cGy). The fraction of the left lung volume that receives 18 or 20 Gy differed significantly for the AI segmentations (24 ± 4 %) and the clinical segmentations (24 ± 4 %). For the right lung, where the volumes that receive 18 or 20 Gy were close to zero, there was no significant difference. Boxplots visualizing these results are shown in Figure 28.

Table 4: Result from the Wilcoxon signed rank test of the differences in dose to the lungs calculated for the clinical segmentations and AI-generated segmentations. Whether V18 or V20 for the lungs is relevant depends on the dose fractionation.

	Paired differences (clinical - AI)			
	Median	Min	Max	p-value
Average left lung dose (cGy)	-8,9	-23,9	48,4	**0,022**
Average right lung dose (cGy)	-0,5	-1,6	0,8	**0,038**
V18/20 left lung (%)	-0,2	-0,01	0,01	**0,016**
V18/20 right lung (%)	0,0	0,0	0,0	0,286



(a) Dose

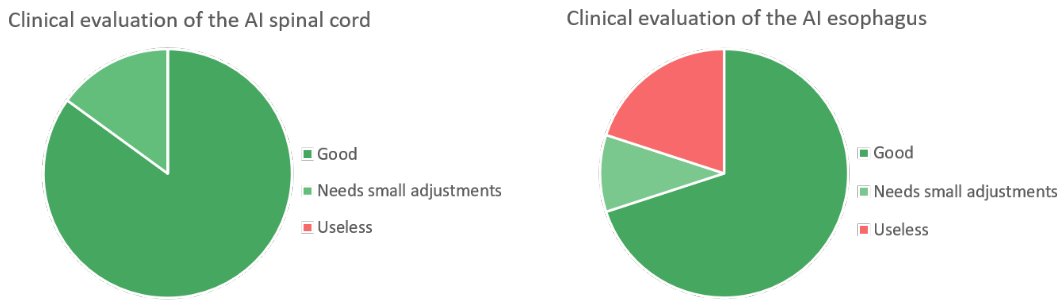
(b) Dose difference

Figure 28: Boxplots with the average lung doses and average lung dose differences (clinical - AI) calculated for the clinical segmentations and the AI-generated segmentations obtained with the DL thorax model.

4.1.3 Spinal cord and esophagus

The result from the clinical evaluation of the AI spinal cord and esophagus segmentations produced with the DL thorax model is shown in Figure 29. Overall, the result was satisfactory. The spinal cord segmentations were deemed good as they are in 85 % of the cases, and 25 % of the cases needed only small adjustments. In general, the AI structures were inaccurate in the top and bottom slices for the spinal cord.

For the esophagus, 70 % of the segmentations were deemed good as they are, 10 % needed only small adjustments, while 20 % were deemed useless. The caudal part of the esophagus was in particular of poor quality in some cases.

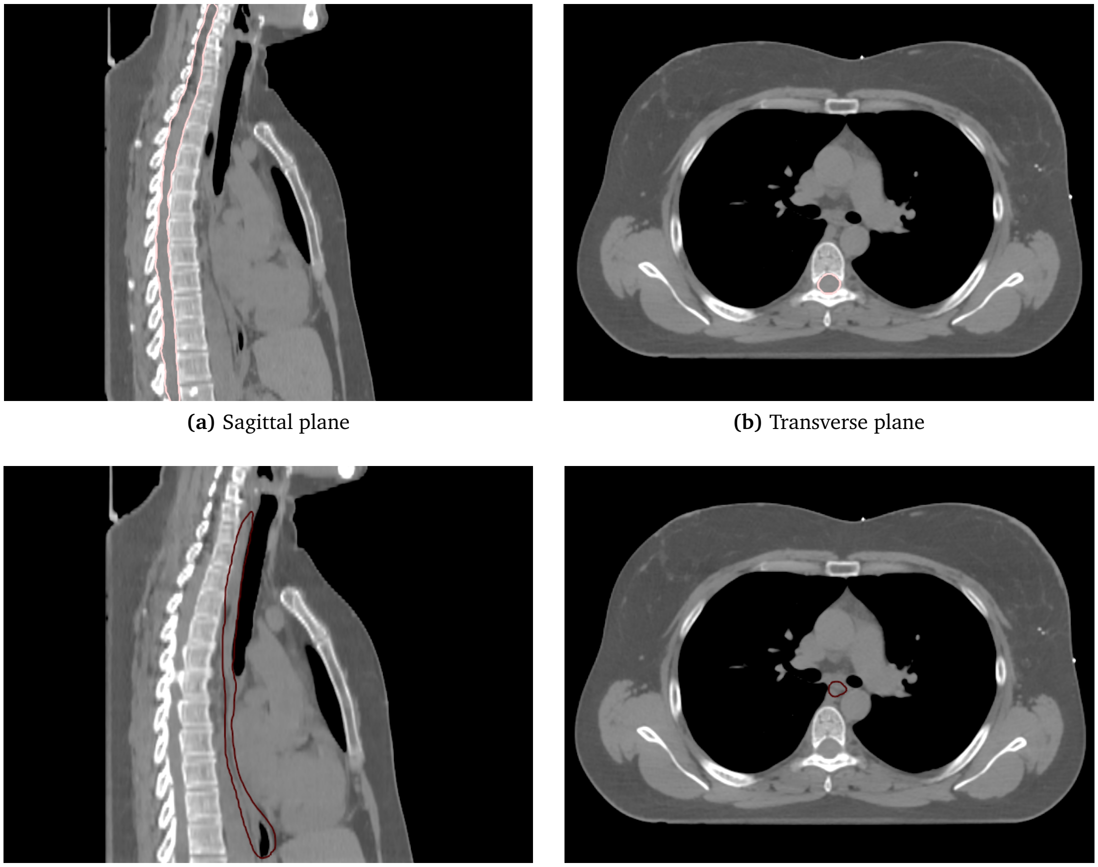


(a) Spinal cord

(b) Esophagus

Figure 29: Clinical assessment of the spinal cord and esophagus segmentations obtained with the DL thorax model. Dark green indicates that the AI segmentation is good as it is; light green suggests that small adjustments are needed, but the structure serves as a good starting point; red indicates that the structure is useless, and starting over again is preferable.

Figure 30 shows a representative example of the esophagus and spinal cord segmentations generated with the DL thorax model for one patient. In this case, both AI segmentations were assessed good as they are.



(a) Sagittal plane

(b) Transverse plane

(c) Sagittal plane

(d) Transverse plane

Figure 30: Example showing the spinal cord (□) and the esophagus (■) segmentations produced with the DL thorax model for one patient.

4.2 Training and testing of ML models

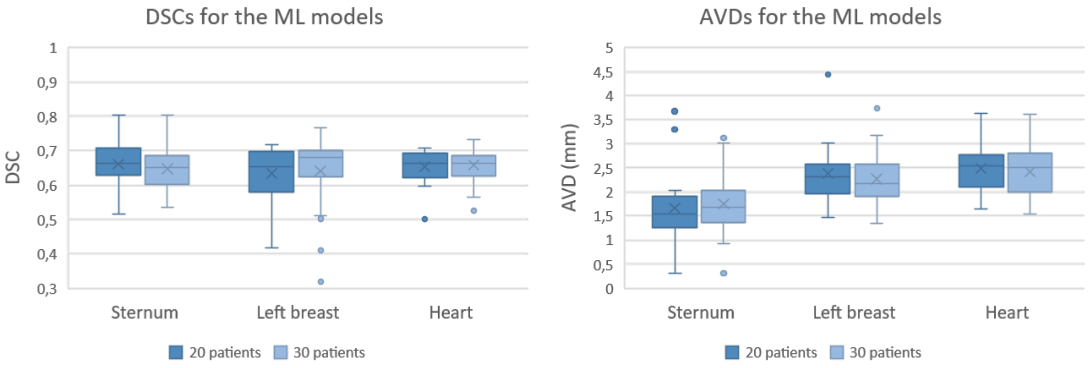
The runtime for the ML models was from 30 seconds to 5 minutes depending on the ROI and the number of patients used for training the model. This time includes loading the data, training the model, and calculating evaluation metrics.

Model performance with different amount of training data

Table 5 and Figure 31 compare the result achieved with the different ML models trained with 20 image series and 30 image series. The result of the Wilcoxon rank sum test is shown in Table 5. Training the ML models with 10 more patients did not affect the result significantly, although the boxplots in Figure 31 may indicate that the result is somewhat worse for the sternum, while small improvements are seen for the left breast and the heart.

Table 5: Result from the Wilcoxon rank sum test of the DSCs and HDs for the ML models trained with 20 and 30 patients.

	Median DSC			Median AVD (mm)		
	20	30	p-value	20	30	p-value
Sternum	0,66	0,65	0,368	1,55	1,68	0,389
Left breast	0,65	0,68	0,533	2,31	2,16	0,572
Heart	0,66	0,66	0,714	2,54	2,51	0,774



(a) DSC

(b) AVD

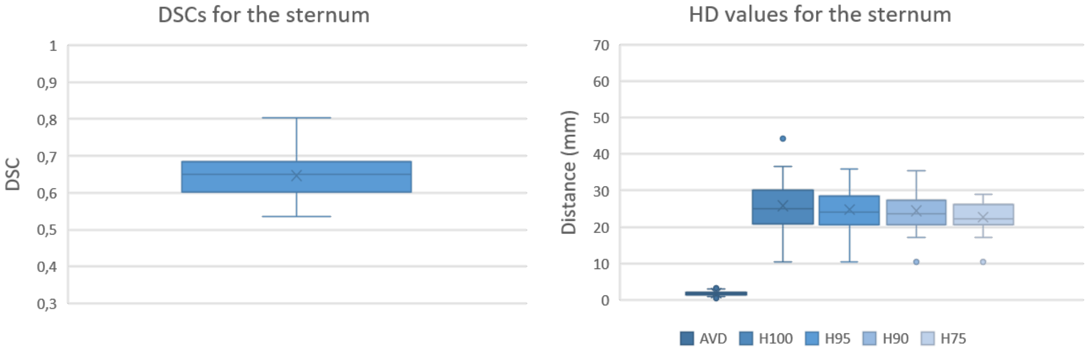
Figure 31: Boxplots with the DSC and AVD results for the ML models trained for the sternum, the left breast, and the heart with 20 and 30 patients.

4.2.1 Sternum

The average values for all metrics for the AI segmentations and the clinical segmentations for the ML model trained for the sternum with 30 image series are displayed in Table 6. The performance of the ML model varied between the individual patients in the dataset, as seen in Figure 32. The lowest and highest DSCs were 0,54 and 0,80, respectively; the lowest and highest AVD values were 0,3 mm and 3,1 mm, respectively.

Table 6: Average values for all metrics for the sternum segmentations obtained with the ML model trained with 30 patients. All HD values are in mm.

Metric	Average	STD	Min	Max
DSC	0,65	0,06	0,54	0,80
H75	22,6	4,1	10,4	28,9
H90	24,4	5,4	10,4	35,5
H95	24,8	5,7	10,4	35,9
H100	25,9	6,9	10,4	44,3
AVD	1,7	0,6	0,3	3,1

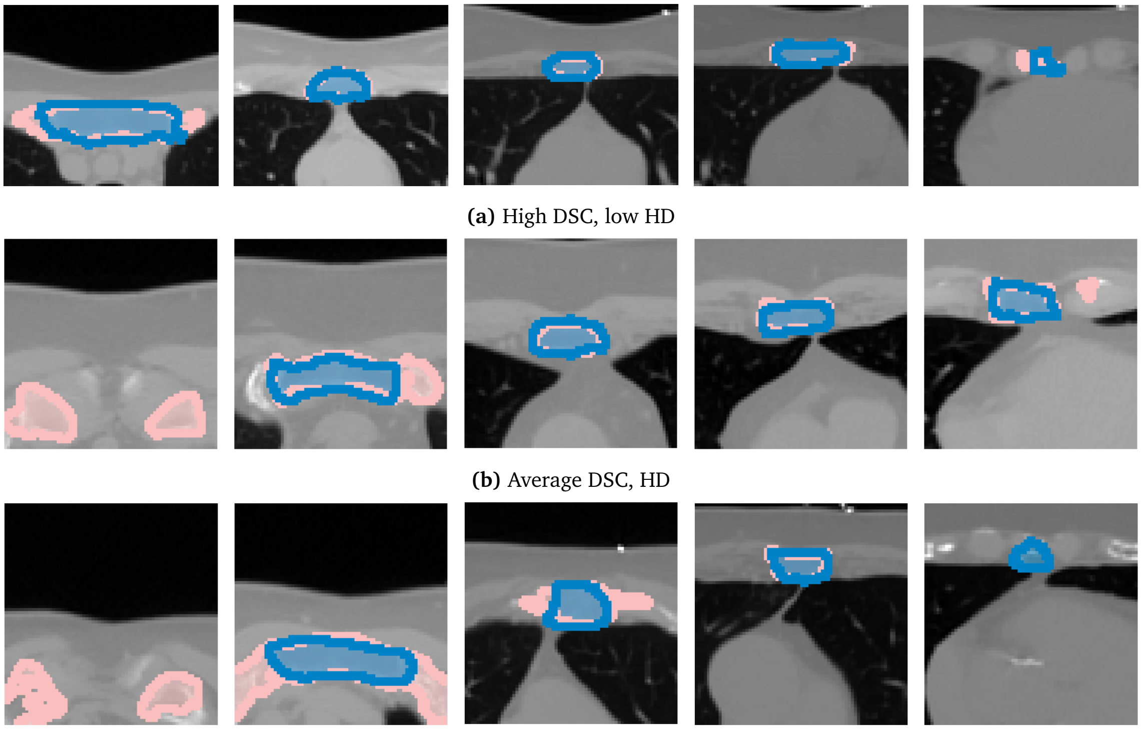


(a) DSC

(b) HD values

Figure 32: Boxplots with the DSC and the different HD results for the AI segmentations obtained with the ML model trained for the sternum with 30 patients.

Figure 33 shows some of the segmentations achieved with the ML model trained for the sternum with 30 image series, together with the manual segmentations. Figure 33a shows a patient that got a high DSC and low HDs, Figure 33b is an example from a patient with a DSC and HDs close to the average performance of the model, and Figure 33c is an example from a patient with a low DSC and high HDs. In general, the model struggled with the first and last slices, and this was seen for every patient. The AI segmentation still maintained a good agreement with the manual segmentation for the middle slices.



(a) High DSC, low HD

(b) Average DSC, HD

(c) Low DSC, high HD

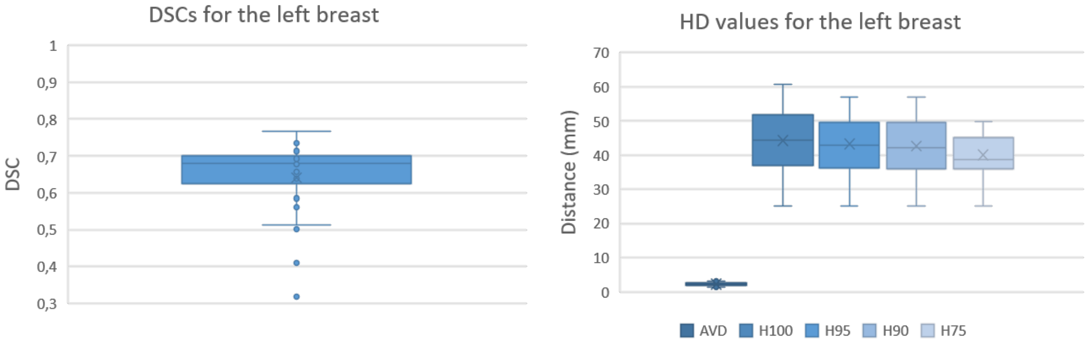
Figure 33: Examples from three different patients showing the sternum segmentations produced with the ML model (🟥) and the manual sternum segmentations (🟦). From left to right: cranial to caudal.

4.2.2 Left breast

The average values for all metrics for the AI segmentations and the clinical segmentations for the ML model trained for the left breast with 30 image series are displayed in Table 7. The model performed less uniform in terms of DSC for the left breast compared to for the sternum, as seen in Figure 34. The lowest and highest DSCs were 0,32 and 0,77, respectively; the lowest and highest AVD values were 1,3 mm and 3,7 mm, respectively.

Table 7: Average values for all metrics for the left breast segmentations obtained with the ML model trained with 30 patients. All HD values are in mm.

Metric	Average	STD	Min	Max
DSC	0,64	0,10	0,32	0,77
H75	40,0	6,5	25,0	49,7
H90	42,7	8,2	25,0	56,9
H95	43,2	8,5	25,0	57,0
H100	44,3	9,1	25,0	60,7
AVD	2,3	0,5	1,3	3,7

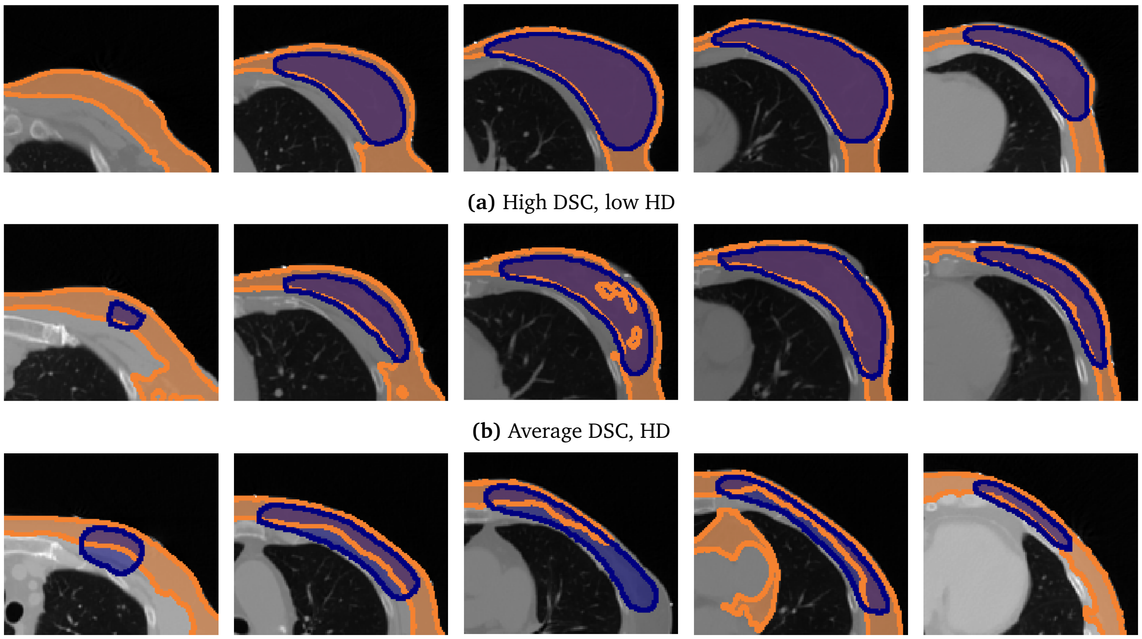


(a) DSC

(b) HD values

Figure 34: Boxplots with the DSC and the different HD results for the AI segmentations obtained with the ML model trained for the left breast with 30 patients.

Figure 35 shows some of the segmentations achieved with the ML model trained for the left breast with 30 image series, together with the clinical segmentations. Figure 35a shows a patient that got a high DSC and low HDs, Figure 35b is an example from a patient with a DSC and HDs close to the average performance of the model, and Figure 35c shows a case with a low DSC and high HDs. Also in this case, the model struggled especially with the first and last slices.



(a) High DSC, low HD

(b) Average DSC, HD

(c) Low DSC, high HD

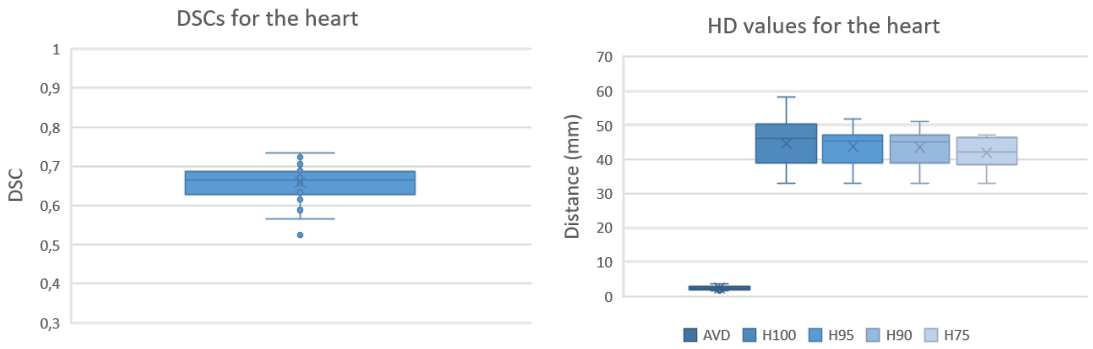
Figure 35: Examples from three different patients showing the left breast segmentations produced with the ML model (🟧) and the clinical breast segmentations (🟦). From left to right: cranial to caudal.

4.2.3 Heart

Table 8 shows the average values for all metrics for the AI segmentations and the clinical segmentations for the ML model trained for the heart with 30 image series. The ML model for the heart performed better and more uniform in terms of DSC compared to for the sternum and the left breast. Figure 36 shows how the model varied between the patients. The lowest and highest DSCs were 0,52 and 0,73, respectively; the lowest and highest AVD values were 1,5 mm and 3,6 mm, respectively.

Table 8: Average values for all metrics for the heart segmentations obtained with the ML model trained with 30 patients. All HD values are in mm.

Metric	Average	STD	Min	Max
DSC	0,66	0,05	0,52	0,73
H75	41,9	4,4	33,0	47,0
H90	43,5	5,2	33,0	51,0
H95	43,8	5,3	33,0	51,7
H100	44,6	6,0	33,0	58,3
AVD	2,4	0,5	1,5	3,6

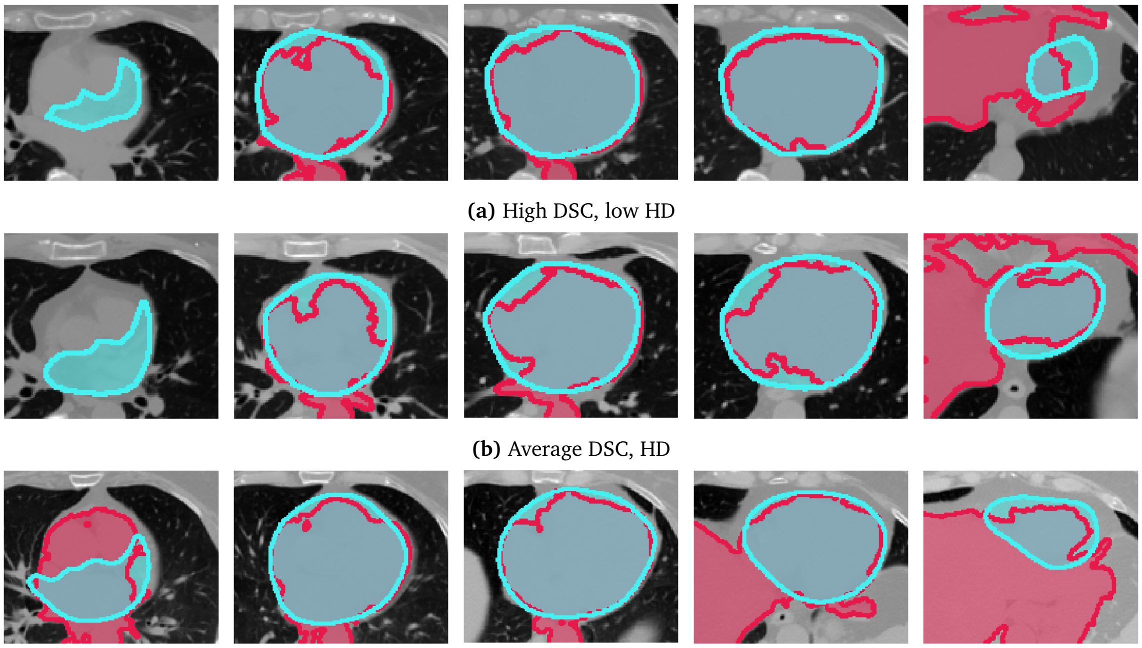


(a) DSC

(b) HD values

Figure 36: Boxplots with the DSC and the different HD results for the AI segmentations obtained with the ML model trained for the heart with 30 patients.

Figure 37 shows some of the segmentations achieved with the ML model trained for the heart with 30 patients, together with the clinical segmentations. Figure 37a shows a case with a high DSC and low HDs, Figure 37b is an example from a patient with a DSC and HDs close to the average performance of the model, and Figure 37c shows a patient with a low DSC and high HDs. The model struggled especially with the first and last slices for the heart as well, while the middle slices of the AI segmentations maintained a better agreement with the clinical segmentations.



(a) High DSC, low HD

(b) Average DSC, HD

(c) Low DSC, high HD

Figure 37: Examples from three different patients showing the heart segmentations produced with the ML model (🟥) and the clinical heart segmentations (🟦). From left to right: cranial to caudal.

5 Discussion

5.1 Evaluation of a DL thorax model

A DL model for automatic segmentation of organs in the thorax region, implemented in a commercial treatment planning system, was evaluated. The AI segmentations of the heart and the lungs were evaluated with comparison to the clinical segmentations. The average DSC and AVD for the heart were 0,92 ± 0,02 and 2,9 ± 1,1 mm, respectively. For the lungs, the average DSC and AVD were 0,97 ± 0,01 and 0,9 ± 0,4 mm, respectively. Although there is no consensus on the interpretation of overlap indices, a DSC greater than 0,7 is commonly used to indicate good agreement [45, 46, 47]. DL approaches have shown promising results in thoracic segmentation. Lei et al. [48] reported an average DSC of 0,87, 0,97, 0,90, and 0,75 for the heart, lungs, spinal cord, and esophagus, respectively. Mamani et al. [49] reported an average DSC of 0,95 for the lungs, and Trullo et al. [50] obtained an average DSC of 0,67 for the esophagus and 0,90 for the heart. Compared to these results, the DL thorax model performed especially well, in terms of DSC, for the heart and lung segmentations.

Additionally, five different DL-based methods for auto-segmentation were developed by different institutes for the thoracic auto-segmentation challenge organized at the 2017 annual meeting of American Association of Physicists in Medicine (AAPM) [51]. The DSC and AVD segmentation results from the five institutes are listed together with the results for the DL thorax model in Table 9. The average DSC ranged between 0,85 and 0,93 for the heart, 0,95 and 0,98 for the lungs, 0,83 and 0,89 for the spinal cord, and 0,55 and 0,72 for the esophagus. The DL thorax model achieved similar DSC and AVD for the heart and the lungs.

These findings also indicate that it is difficult to achieve as good results for the esophagus as for the other OARs. This can be explained by two factors that are important for segmentation performance: the visualization of the boundary of the organ and the organ volume. For instance, the lungs have high contrast edges that are relatively easy to detect for both software and a human observer. Contrarily, the esophagus has low contrast edges, which are much harder to detect. As for manual segmentation, auto-segmentation methods are therefore typically less accurate for small, less visible soft-tissue boundaries, such as the esophagus. Although the esophagus segmentation with the DL thorax model was satisfactory, 4 out of the 20 the cases were deemed useless by the reviewer. The addition of multi-modal images, such as MR imaging that provide improved soft tissue contrast, could potentially improve the result for the esophagus.

Table 9: A segmentation comparison of the DL thorax model with the five DL methods that participated in the AAPM thoracic auto-segmentation challenge [51]. The results are given in average DSC and AVD, where bold indicates the best values.

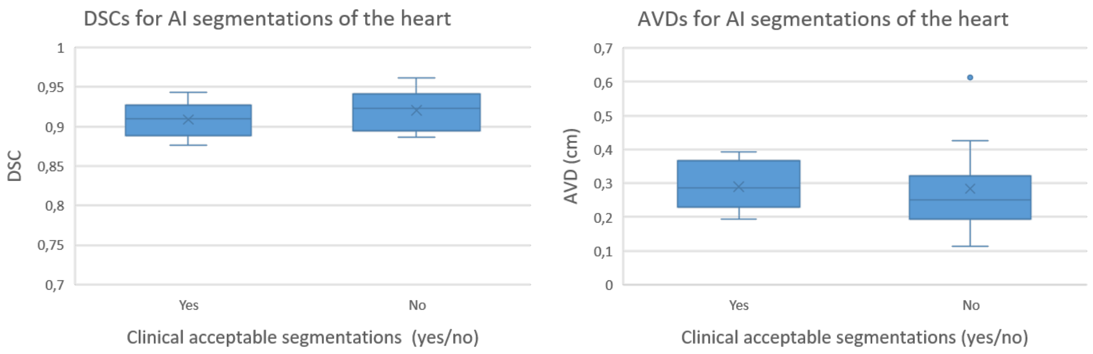
DSC	Heart	Left lung	Right lung	Spinal cord	Esophagus
Model 1	**0,93 ± 0,02**	0,97 ± 0,02	0,97 ± 0,02	0,88 ± 0,04	**0,72 ± 0,10**
Model 2	0,92 ± 0,02	**0,98 ± 0,01**	0,97 ± 0,02	**0,89 ± 0,04**	0,64 ± 0,20
Model 3	0,91 ± 0,02	0,98 ± 0,02	0,97 ± 0,02	0,87 ± 0,11	0,71 ± 0,12
Model 4	0,92 ± 0,02	0,96 ± 0,03	0,95 ± 0,05	0,85 ± 0,04	0,61 ± 0,11
Model 5	0,85 ± 0,04	0,95 ± 0,03	0,96 ± 0,02	0,83 ± 0,08	0,55 ± 0,20
DL thorax	0,92 ± 0,02	0,97 ± 0,01	**0,98 ± 0,01**		
AVD (mm)					
Model 1	**2,05 ± 0,62**	0,74 ± 0,31	1,08 ± 0,54	0,73 ± 0,21	2,23 ± 2,82
Model 2	2,42 ± 0,82	**0,61 ± 0,26**	0,93 ± 0,53	**0,69 ± 0,25**	6,30 ± 9,08
Model 3	2,98 ± 0,93	0,62 ± 0,35	0,91 ± 0,52	0,76 ± 0,60	**2,08 ± 1,94**
Model 4	2,61 ± 0,69	2,90 ± 6,94	2,70 ± 4,84	1,03 ± 0,84	2,48 ± 1,15
Model 5	4,55 ± 1,59	1,22 ± 0,61	1,13 ± 0,49	2,10 ± 2,49	13,1 ± 10,4
DL thorax	2,87 ± 1,11	1,06 ± 0,04	**0,79 ± 0,03**		

The dosimetric impact of the segmentations obtained with the DL thorax model was also evaluated by comparing the dose to the heart and the lungs for the clinical and AI-generated segmentations. The average heart doses did not differ significantly, while for the average lung doses and the fraction of the left lung volume that receives 18 or 20 Gy, the p-values were less than 0,05, which indicates a statistically significant dose difference. The left lung dose metrics for both clinical and AI-generated segmentations were all less than 14,17 Gy, and the dose differences were less than 0,24 Gy, except for one patient with a dose difference of 0,48 Gy for average dose. For the right lung, the dose metrics for both clinical and AI-generated segmentations were all less than 2,17 Gy, and the dose differences were less than 0,02 Gy. Although the p-values were less than 0,05, the dose differences of 0,00 Gy to 0,48 Gy are minimal.

Whether differences between clinical and AI-generated segmentations result in clinically relevant alterations in calculated doses to OARs depends partly on proximity of normal structures to the treatment volume and the dose gradient. For all patients, the left lung lies closest to the target volume, and it is therefore expected that the dose metrics and dose metric differences in the left lung are larger than those of the heart and right lung. Even though the average heart dose differences were not significant, the largest individual differences were seen for the heart. For 6 of the 20 patients, the average heart dose differences were more than 0,30 Gy. In clinical evaluation, the treatment planner strives for an average heart dose below 2 Gy, which means that a difference of 0,5 Gy may change the treatment plan.

Although the DSC and HD have shown to be good measures for geometric similarity, they do not always correlate with clinical applicability of the segmentations. In clinical practice, segmentations are either made by or reviewed by a physician. The AI segmentations made with the DL thorax model were therefore reviewed qualitatively by an experienced and skilled oncologist at St. Olavs Hospital. The clinical approval was not based on any predefined criteria but was a qualitative evaluation made by the reviewer for each ROI and patient. The AI-generated heart segmentations passed in 42 % of the cases, in terms of clinical acceptability. For the lungs, all of the AI-generated segmentations were assessed clinically acceptable, although the clinical segmentations were preferred over the AI segmentations in 70 % of the cases. The spinal cord and esophagus segmentations were satisfactory and only 10 % of the segmentations were deemed useless. Especially for the heart, where the AI segmentations were not approved in 12 of the 20 cases, it would be useful to know how far away the non-approved segmentations were from being approved. A way for scoring the non-approved segmentations could therefore be beneficial. For example, each segmentation could be given a score based on how demanding or time-consuming it is to adjust in order to make it acceptable according to local clinical standards. A method similar to this was conducted by Lusting et al. [29] for example.

Figure 38 shows boxplots of the DSCs and AVDs grouped for the clinical accepted and non-accepted heart segmentations for the DL thorax model. There was no significant difference between the mean of the DSCs for the two groups (p = 0,279), nor for the mean of the AVDs for the two groups (p = 0,897).



(a) DSC

(b) AVD

Figure 38: Boxplots where DSCs and AVDs for segmentations of the heart for the DL thorax model are grouped by clinical and non-clinical acceptable segmentations.

An important point in this analysis is that the clinical segmentations were not critically reviewed previous to this study, and they were contoured by several different physicians. But after all, they were used for treatment and should therefore be of good quality. However, the comments from the reviewer suggest that there was certainly room for improvements in the manual segmentations as well. The ground truth that the AI-generated structures were compared to should therefore have been better quality assured to make the quantitative evaluation more valuable.

The DL thorax model used on average about 3 minutes on generating AI segmentations for one patient. Even though 21 % of the AI segmentations required further improvements or adjustments, using them as a starting point for manual segmentation may save time. Many studies show time-savings compared to full manual segmentation [29, 52, 53, 54, 55].

Lustberg et al. [29] provided a clinical evaluation of a DL-based method for automatic segmentation for radiotherapy treatment planning for lung cancer using commercial software (Mirada Medical Ltd., Oxford, United Kingdom). Similar to the results for the DL thorax model, this DL-based method performed well for the lungs, with DSCs > 0,97, and the segmentations needed little or no corrections to conform to local clinical standards. Likewise, many of the heart segmentations gave satisfactory results, with DSCs > 0,80, while some of the segmentations needed further editing. These researchers showed that time was saved when using the auto-generated segmentations as starting point for manual segmentation: The total median time saved was 10 minutes for the AI segmentations with user adjustment with respect to the manual segmentations. This is a large reduction compared to the median time required to contour the OARs, including the heart, the left and right lung, the esophagus, the spinal cord, and the mediastinum, which for Lustberg et al. were 20 minutes. This means that the median time required to contour all OARs was halved with the DL-based method.

As mentioned, further improvement was especially needed for the heart segmentations produced with the DL thorax model; adjustments were needed in 12 of the 20 cases, according to the reviewer. For further work, it would be interesting to measure how much more time would be needed to make all AI segmentations clinically acceptable. For instance, Schreier et al. [56] constructed a DL method for auto-segmentation of the breasts and heart, where they asked two experienced dosimetrists and two radiation oncology specialists to correct the AI-generated segmentations to make them clinically acceptable according to their guidelines and measure the time needed for the corrections. Using an approach like this would be to show time-savings, if any, for the auto-segmentation, despite the need for manual corrections, compared to full manual segmentation.

The DL thorax model came pre-trained in RayStation and was trained with 65 segmented image sets for lung cancer patients, originating from Centre Oscar Lambret. The model performed well for the breast cancer patients treated with deep inspiration breath hold despite being trained with lung cancer patients. It would still be interesting to see if training the model with local data could improve the results. This would ensure that the segmentation guidelines are in accordance with those used at St. Olavs Hospital and thus the data used to evaluate the model. Additionally, training the model with more image sets could potentially improve the results. It should also be mentioned that the DL thorax model has not been validated by RaySearch.

5.2 Training and testing of ML models

An ML method using linear SVC for automatic segmentation in medical images, developed at NTNU, was trained for segmentation of the sternum, the left breast, and the heart and validated in terms of accuracy. Training the ML models with 20 or 30 patients made no significant difference to the results in study. This was possibly due to a large variation in the dataset. However, the models trained with 30 patients were considered for further analysis, because more training data should in theory improve the result. In general, the models struggled with the first and last slices but performed better for the middle slices. For the ML models trained with 30 images, the average DSC and AVD were 0,65 ± 0,06 and 1,8 ± 0,6 mm, respectively, for sternum. For the breast, the average DSC and AVD were 0,64 ± 0,10 and 2,3 ± 0,5 mm, respectively, and for the heart, the average DSC and AVD were 0,65 ± 0,05 and 2,4 ± 0,5 mm, respectively. The planning of conformal radiotherapy requires accurate segmentations of ROIs and this result is not good enough for clinical use.

The predicted volumes for most of the structures in the dataset contained a large amount of false positives. This means that relatively large areas with non-organ voxels were predicted as the organ. This was seen for all three ROIs but especially for the left breast, the first slices of the sternum, and the last slices of the heart. The patient that got one of the lowest DSCs for the breast segmentation, shown in Figure 35c, had a relatively small left breast compared to the other patients. It is reasonable to think that with a high amount of false positives, the model will perform worse in terms of DSC for small structures. For this patient, the DSC would be relatively low even if the whole breast volume was classified correct. This means that the DSC will be very sensitive to the number of false positives and false negatives compared to the DSC for a patient with a larger breast. This property of the DSC is further discussed in section 5.3.

Post-processing was an important step in the development of the model. Especially, the removal of small areas and volumes from the predicted volumes increased the average DSC. The threshold for the size was adjusted to give the best results overall; however, for some patients it looked like this threshold was set too high, and the actual ROI volume was removed. This was seen for some patients where the AI structure of the sternum was missing in the last slices. Also, the large volumes that the model incorrectly predicted as the ROI have a much larger impact on the results compared to the small volumes. Removing these would have a larger effect on the results, but it would also be more challenging.

One method that could have been used to remove these larger incorrect volumes is a click approach that simulates the physician clicking on the selected structure, and everything that is not connected to the structure is removed. This method separates connected regions and gets a list of seeds, which are positions that are inside the ground truth mask, before removing all objects not connected to a seed. This method could be effective to remove the large number of false positives seen for many of the predictions and in that way improve the result. Such a method is possible to implement with functions from the SITK library for example.

It should be noted that all patients had lead wires placed on the skin around the palpable breast. This wire is placed on the patient by a physician prior to CT scanning and is used to help defining the breast for when it is segmented afterwards. It is reasonable to think that this lead wire may have had an impact on the way the ML model for the left breast was thinking. It is conceivable that the wire is useful as it helps to define the breast. However, the images in Figure 35 shows that it is more likely that the wire has confused the algorithm rather than helping it, if it has affected the model at all.

While few auto-segmentation methods using SVC for thoracic OARs have been published, support vector machines have shown promising results for segmentation of brain tumors [57] and tumors in the prostate [58] with MR imaging. Dong et al. [59] have suggested an ML method using random forest classification for efficient mass segmentation for breast cancer patients. The model was compared to different support vector machines, and the proposed random forest classifier outperformed all of the other methods. This suggests that random forest classification may be superior compared to a support vector classifier in tumor segmentation for breast cancer patients. Yet, it is unknown if these results apply to segmentation of the OARs.

One advantage of the ML algorithm is that it is very flexible, and it can easily be extended to include other ROIs or images types. However, the amount of training data applied could potentially have limited the performance of the models. It could be that the models do not generalize well after training, and therefore perform poorly on new, unseen data. An attempt to improve the models was conducted by training them with 10 more patients; however, this did not result in any significant change to the results. Augmented data could also have been used to increase the training data. Yet, it is not given that augmented data can represent realistic subjects and increase the generalization of a model. An even larger dataset could potentially provide results that are more realistic.

There is no definite answer to what amount of data is needed for training, but the amount of data required is generally less for ML approaches compared to DL approaches, because the learning algorithm is less complex. This is an advantage of ML as high quality data is often hard to collect due to patient data privacy and the fact that an absolute ground truth does not exist. Also, very large imaging datasets will require large storage and memory requirements along with high training time for the models. However, DL algorithms often perform better when given more data, which is not typical for ML algorithms after a plateau is reached. Yet, it would be interesting to see if the ML models would perform better when trained on more patients.

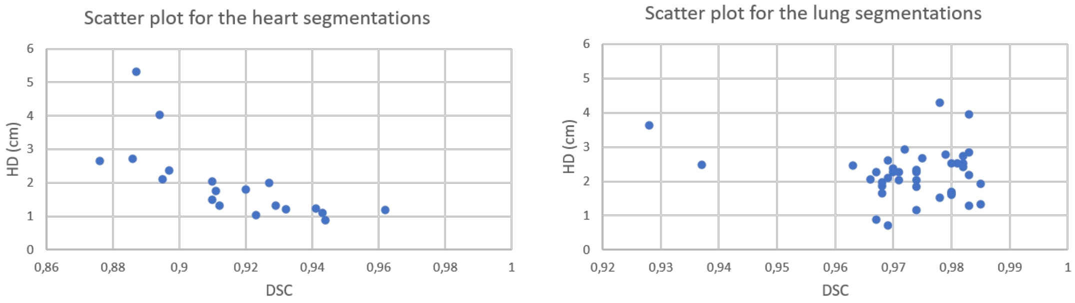
Also, it is important to ensure the quality of the data put into the model. The ROI segmentations used as ground truth were not reviewed thoroughly prior to training the model. It would have been favorable if all structures were segmented by the same person to reduce interobserver variation. Another option that could be explored is to use the union or the intersection of segmentations from two different experts as ground truth.

5.3 Metrics used for comparison

While many methods are available for comparing segmentations in contouring studies [4], clinically relevant segmentation evaluation remains challenging. Selecting suitable metrics is not a trivial task as metrics have different properties, with biases and sensitivities. Overlap methods account for both volume and positional variability between segmentations and are widely used [4, 60]. The most popular overlap method is the DSC; Sharp et al. [2] stated in their review on auto-segmentation methods that this index should be included in any evaluation as it is the most commonly reported metric used in literature. Thus, the model performances were evaluated in terms of the DSC for the predicted segmentations and the clinical segmentations in the present study.

Spatial distance based metrics are widely used in the evaluation of image segmentation as dissimilarity measures. They are recommended when the overall accuracy of the segmentation is of importance [42]. The HD is a distance based metric and was used as a second quantitative evaluation metric to complement the DSC. H100 is generally sensitive to outliers and because noise and outliers are common in medical segmentation, more robust variants of the HD were used as supplements to the H100. This included the quantile method proposed by Huttenlocher et al. [43] and the AVD, which are known to be more stable and less sensitive to outliers compared to the H100. Using percentiles rather than the maximum distance is more robust as issues with noisy segmentations are avoided. In some cases, the segmentations can look good qualitatively but have a few stray voxels. Using the maximum penalizes these cases heavily. So, the H100 is this not a good approach for such cases.

Different metrics may complement each other and do not necessarily correlate. The DSCs and H100s plotted against each other for the DL thorax model is shown in Figure 39. These plots show that there is a weak correlation between the DSC and H100 for the heart, while there is no correlation between the DSC and H100 for the lungs. For the lungs, this means that optimizing for one metric does not optimize the other. The correlation between the DSC and HD generally decreases with decreasing overlap. This is because the DSC, in contrast to the HD, do not consider the position of false positive voxels. This means that the DSC does not consider the positions of voxels that are not in the overlap region and thus provides the same value independent of the distance between the voxels. For the lungs, the predicted AI segmentations have some regions that are not in the overlap, as seen in Figure 27a. These regions will not be considered differently by the DSC but will give different results for the H100. This may explain the lack of correlation between the DSCs and H100s for the lung segmentations.



(a) Heart

(b) Lungs

Figure 39: The DSC and H100 values plotted against each other for the heart and lung segmentations for the DL thorax model.

There is an inverse relation between structure size and the overlap between the structure in the ground truth and that in the segmentation under test. For small structures, the probability of small or zero overlap is high. This is seen in the last slices of some of the sternum segmentations for example, as shown in Figure 40. In such a case, overlap metrics are not suitable, since they providee the same value regardless of how far the structures are from each other, once the overlap is zero. Thus, the HD may be better suited over the DSC as an evaluation metric when structures are small.

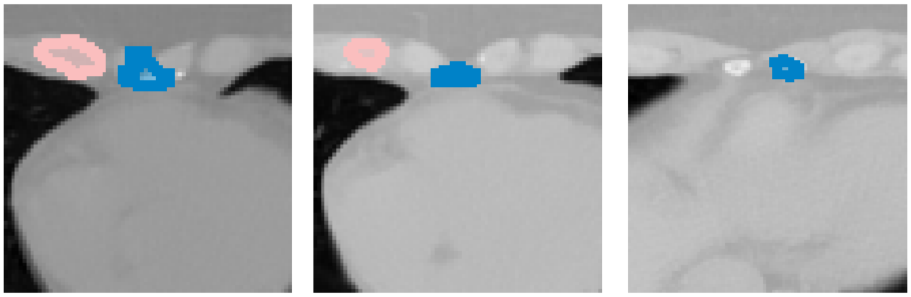


Figure 40: Example showing three different slices for a patient where the DSC for the AI segmentation and the manual segmentation of the sternum is the same and equal zero.

The DSC is correlated with the size of the structure and is therefore organ-dependent. What is assessed good will depend on the clinical context, and the DSC thus become a less meaningful measure when comparing the quality of the segmentations between different organs. It usually achieves higher scores for bigger organs. This can be explained by looking at the definition of the DSC given in Equation (3.2). For small structures, the number of true positives in the prediction will be low even if the whole ROI volume is classified correct. The DSC will then be very sensitive to the number of false positives and false negatives compared to the DSC for a patient with a larger structure. As mentioned, the predictions made by the ML models included a large number of false positives, and the DSC is therefore highly correlated with the size of the structure for these predictions.

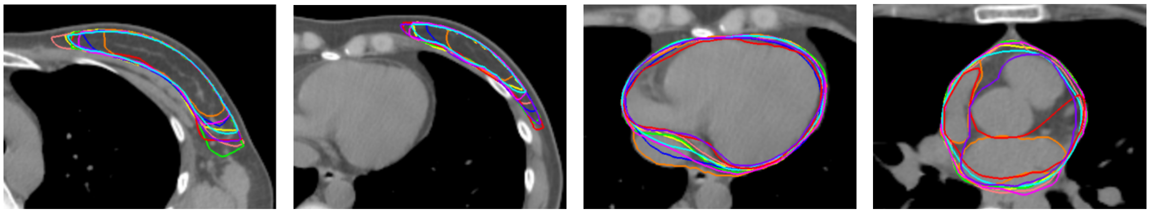
This property may also explain the high DSCs obtained for the lung segmentations with the DL thorax model. It is reasonable to believe that the lung volumes could be so large that the DSC gives a high value even if the obtained AI segmentation differ quite much from the clinical segmentation. Therefore, the HD may be a better choice for comparison of the lung segmentations; although HD is not perfect either, it only says something about the largest deviation.

As mentioned, these metrics are intuitive and quantitative, but they do not always correlate with clinical applicability of the segmentations. Vaassen et al. [40] suggest that quantitative measures to predict time-saving using automatically generated segmentations are better indicators of clinical applicability and quality. These researchers introduced two new evaluation measures: the surface DSC and the added path length (APL). These measures were found to be better indicators for clinical segmentation time saved, and thus clinical applicability and quality, compared to the commonly-used volumetric DSC and HD. For further segmentation studies, these measure could be included for more clinically relevant measures.

Lustberg et al. [29] grouped the DSCs, HDs, and time saved based on how well the OARs performed according to the subjective score of a technician and found that there was a relation between the quantitative measures and the subjective score. This demonstrates that the DSC and HD can correlate with clinical applicability well, even though better alternatives, such as the surface DSC and APL, may exist. Nonetheless, both the DSC and HD are useful for comparison to the works done by other researchers.

Even if the evaluation measures were more correlated with clinical applicability of the segmentations, evaluating knowledge-based segmentation methods still remains challenging, the main reason being the absence of an absolute ground truth that can be directly derived from CT data. In this study, all segmentations are assessed based on deviations from single observer manual segmentations. The segmentations, provided by different observers, are assumed to be correct and are therefore used as a ground truth. However, manual segmentation is prone to intra- and interobserver variations, and quite different segmentations of the same ROI can be accepted in compliance with local clinical guidelines. For instance, Lustberg et al. [29] found that the even after adjusting the automatically generated esophagus segmentation to meet the clinical guidelines, the DSC was 0,78 when comparing to the manual segmentation.

In fact, results presented at the Norwegian Radiotherapy Meeting 2018 [61] showed that target volume and OAR segmentation vary for breast cancer between the different radiotherapy departments in Norway. Prior to the meeting, the different radiotherapy departments received a patient case, where they segmented the target volumes and OARs based on their local guidelines. Figure 41 shows some of the submitted results that were collected and compared for the clinical target volume breast and the heart. The volume of the breast ranged from approximately 320 cm^3 to 360 cm^3 with the different segmentations. For the heart, the volume ranged from approximately 440 cm^3 to 560 cm^3. One reason for these variations may be explained by the different guidelines employed. Also, the segmentation of the heart is performed by different professions at the different radiotherapy departments. In half of the departments, physicians segment the heart; in the other half, radiation therapists segment the heart, while a physician control the segmentation afterwards.



(a) Left breast

(b) Heart

Figure 41: The left breast and the heart manually segmented for the same patient at different radiotherapy departments in Norway — courtesy of [61].

Again, the ground truths that the AI-generated structures were compared to should have been reviewed prior to this study: Better quality assurance of the ground truths would have made the quantitative evaluation more meaningful. For further work, it could also be interesting to compare the DSC and HD result using ground truth segmentations from different experts. This would be to measure how much the quantitative result is affected by using another set of ground truths. In addition, one could see if the accuracy of the AI-generated segmentations are similar to expert inter-observer variability. For instance, Wong et al. [62] proposed a DL method for auto-segmentation of OARs and compared the differences in DSC and H95 for different expert segmentations and for the DL method and the expert segmentations.

5.4 Management of patient data

The use of AI in radiotherapy challenges in many areas, and particularly is the question of privacy and security of medical data. Using AI on sensitive, personal data for training ML or DL models could be problematic as it may violate the health data law [63].

Regarding the management of patient data for evaluating the DL thorax model in RayStation, all patient data was fully anonymized. For each patient, the original planning CT, together with the segmented target volumes and OARs were used. This information was anonymized in the clinical version of RayStation 8, meaning that all patient information, including name, social security number, date of birth, and gender, was permanently removed from the data. In addition, the date, time, and institution name were removed. The fully anonymized planning CT, together with the segmented target volumes and OARs, were then exported to a non-clinical installation of RayStation 9A.

As for the data used to train the ML models implemented in Python, this patient data was also anonymized in the same manner before being exported. Once the model is trained, the model itself does not contain any patient-identifiable information nor medical images from the training dataset. A fully trained model consists of a single file that is entirely anonymous. As a result, the model can be utilized to run predictions on new, unseen patients independent of sensitive medical data. Also, the use of this data was applied for and approved by the Regional Committees for Medical and Health Research Ethics (REK Midt ref. 92685). REK considered that the project was sound and that the interests of the participants' welfare and integrity would be taken care of. Exemption from the duty of confidentiality were therefore given so that CT images and contoured structures could be de-idenitfied without breaching confidentiality. The committee justified this decision with the fact that the exposed information is CT images and structures, and this health information will be anonymous on the researcher's hand. Additionally, the research project was considered to serve the interests of society, and the risk and privacy disadvantage of participating in the study was considered to be minimal.

Further, there is the challenge of creation and curation of large datasets. Although it is unlikely that robust models can be built with data from a single institution alone, barriers associated with the use of patient data can be a substantial challenge to the development of such models [8]. ML and DL models require patient data, and model training relies on access to large datasets of high quality data. However, obtaining sufficient data is a challenge as labeling image data requires expert knowledge. Collaboration between multiple hospitals could address this challenge but is difficult because of strict internal policies for data sharing and privacy protection imposed by most clinical sites. Sheller et al. [64] introduced the use of federated learning. This is a technique that enables collaborating clinical institutions to train ML models without sharing patient data, thus addressing critical issues such as data privacy, data security, and data access rights.

Another potential solution to the challenges associated with data sharing between institutions is the use of distributed learning, a technique that learns from data without the data leaving the hospital. Jochems et al. [65] developed a predictive model based on a large volume of historical patient data and serves as a proof of concept to demonstrate the distributed learning approach. Using this approach, it is the model that is moved, not the data. Sharing of already trained models is not a problem as the trained ML models and DNNs have no patient data in them. This means that a model can be used in another clinic or another country.

Further, establishing a large, publicly available database of clinical cases with ground truth estimated from multiple expert segmentations would enable standardization of evaluation criteria and provide training data for development of ML algorithms. A successful first step in this direction is the publicly available clinical data provided by The Cancer Imaging Archive [66], a service which de-identifies and hosts a large archive of medical images of cancer patients. However, the way towards availability of large collections of clinical data is still difficult. An absolute ground truth must be decided on, which may be challenging because different clinics have different traditions and standards in their segmentation routines.

There is great potential for better use of resources with less time spent on segmentation and more uniform quality; however, it is currently uncertain how well this potential can be exploited due to challenges related to privacy and data security.

5.5 Further work

As this work included patient CT scans with ROIs segmented by several different physicians, it would be interesting to repeat the comparisons of the AI-generated segmentations and the clinical segmentations with more consistent and critically reviewed clinical segmentations. This would make the quantitative analysis of the segmentation results more meaningful, as well as proving the ML models with training data of higher quality.

It would also be beneficial to extend the clinical analysis of the AI segmentations made by the DL thorax model to provide more information about potential time savings. The ideal would be to measure how much time a physician would spend on manual corrections to make all AI-generated segmentations clinically acceptable. This would be to see if auto-segmentation with AI could save time in total, despite the need for manual adjustments in some of the AI-generated segmentations. However, such a study would require more time and resources. Further improvements of the study could be achieved by comparing the DL thorax segmentation to other auto-segmentation methods implemented in the treatment planning system, such as atlas- or model-based auto-segmentation. For example, the same data could be used for generating atlases in RayStation. Comparing the AI-based segmentations with atlas-based segmentations would be very relevant; maybe some structures could just as easily be based on atlases?

For the ML models, there is large room for improvements. Increasing the amount of training data with 10 patients did not improve the results significantly, but a larger increase in training data may still be of importance and should be further explored. Also, it would be interesting to see if other ML methods, such as random forest classification, could achieve better results than the linear support vector classifier. Specific values used for pre- and post-processing of the image series used as input to the ML models were selected based on a trial and error method. While there are many options for processing the data, only a few were tested. In this study, locally cropping around the ROI and multiple image sets with different window/level were used as input data. For future work, adding new features to the input data, such as texture features or new window/levels, should be further investigated. For post-processing, the mentioned click approach would be interesting to test, as well as more morphological operations to remove holes and smooth boarders, for example. Future studies should also address the pre- and post-processing in a more detailed manner.

The whole dataset was used for the cross-validation, and how the ML models perform on previously unseen data was therefore not explored. Future work should consider to also test the models on a different dataset to see how well the they generalize.

Further, it would be of interest to extend the ML models to include more structures, such as the thyroid, caput humeri, esophagus, and trachea. Especially, anatomically well-defined normal structures, such as caput humeri, could have potential to give good results. It would also have been interesting to train the algorithm for other diagnoses than breast cancer. The ML algorithm has only been used for automatic segmentation of tumor volume in rectal cancer before. For instance, an ML model could be trained for segmentation of brain tumors or tumors in the prostate.

In principle, it should be possible to upload the ML models in RayStation and use them directly as a script to contour ROIs. With this comes a potential for improving the model for every new patient. For further work, it would be of high interest to implement the ML models in RayStation to see how well they work directly in the treatment planning system. The models could then be trained for any structure available, and training data would be easily accessible without having to leave the clinic.

6 Conclusion

A DL thorax model was evaluated for radiotherapy planning for 20 left-sided breast cancer patients using commercial software. Auto-segmentation with this model provided segmentations of high quality, with an average DSC and AVD of 0,92 ± 0,02 and 2,9 ± 1,1 mm, respectively, for the heart and an average DSC and AVD of 0,97 ± 0,01 and 0,9 ± 0,4 mm, respectively, for the lungs. The model generated clinically acceptable results in 42 % of the cases for the heart, 100 % of the cases for the lungs, 85 % of the cases for the spinal cord, and 70 % of the cases for the esophagus. As a large majority of the segmentations were acceptable, and many of the non-accepted segmentations required minor manual corrections, this implies that the model has potential to improve both consistency and efficiency of segmentation in the clinic.

Additionally, ML models for automatic segmentation of the sternum, the left breast, and the heart were trained and tested for 30 left-sided breast cancer patients. The ML algorithm was successfully adapted to train models based on clinical breast cancer segmentations, but the models need further improvements in order to be clinically useful. The average DSC and AVD for the sternum were 0,65 ± 0,06 and 1,8 ± 0,6 mm, respectively; the average DSC and AVD for the left breast were 0,64 ± 0,10 and 2,3 ± 0,5 mm, respectively; the average DSC and AVD for the heart were 0,66 ± 0,05 and 2,4 ± 0,5 mm, respectively. This is a fast and flexible method that can easily be extended to include other anatomical structures or image types. However, there are many options to improve the results, including pre- and post-processing of the data, that should be further explored before the model can be implemented in clinical practice.

To conclude, this study demonstrates that auto-segmentation methods based on AI have potential as a useful tool in radiotherapy planning.

Bibliography

[1] Segedin, B. & Petric, P. 2016. Uncertainties in target volume delineation in radiotherapy - are they relevant and what can we do about them? *Radiology and oncology*, 50(3), 254–262.

[2] Sharp, G., Fritscher, K. D., Pekar, V., Peroni, M., Shusharina, N., Veeraraghavan, H., & Yang, J. 2014. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Medical Physics*, 41(5), 050902.

[3] Schick, K., Sisson, T., Frantzis, J., Khoo, E., & Middleton, M. 2011. An assessment of oar delineation by the radiation therapist. *Radiotherapy and Oncology*, 17(3), 183–187.

[4] Vinod, S. K., Jameson, M. G., Min, M., & Hollowaya, L. C. 2016. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiotherapy and Oncology*, 121(2), 169–179.

[5] Cardenas CE, Yang J, A. B. C. L. B. K. 2019. Advances in auto-segmentation. *Seminars in Radiation Oncology*, 41(5).

[6] Rattan, R., Kataria, T., Banerjee, S., Goyal, S., Gupta, D., Pandita, A., Bisht, S., Narang, K., & Mishra, S. R. 2019. Artificial intelligence in oncology, its scope and future prospects with specific reference to radiation oncology. *BJR|Open*, 1(1), 20180031.

[7] Osman, A. *Radiation Oncology in the Era of Big Data and Machine Learning for Precision Medicine*, 1–30. 03 2019.

[8] Feng, M., Valdes, G., Dixit, N., & Solberg, T. D. 2018. Machine learning in radiation oncology: Opportunities, requirements, and needs. *Frontiers in Oncology*, 8(110).

[9] Mayles, P., Nahum, A., & Rosenwald, J. C. 2007. *Handbook of radiotherapy physics: theory and practice*. Taylor Francis group.

[10] Almberg, S. S. August 2019. Lecture notes in megavoltage electron accelerators for cancer treatment with photons and electrons.

[11] Elith, C., Dempsey, S. E., Findlay, N., & Warren-Forward, H. M. 2011. An introduction to the intensity-modulated radiation therapy (imrt) techniques, tomotherapy, and vmat. *Journal of Medical Imaging and Radiation Sciences*, 42(1), 37–43.

[12] Danielsen, S. September 2019. Lecture notes in treatment planning and treatment techniques.

[13] Helsedirektoratet. 5.4 strålebehandlingsteknikker. URL: https://www.helsebiblioteket.no/retningslinjer/analkreft/kurativ-behandling-av-lokalisert-sykdom/stralebehandlingsteknikker.

[14] Rohlfing, T., Brand, t. R., Menzel, R., Russakoff, D. B., & Maure, r. C. R. 2005. *Handbook of Biomedical Image Analysis*. Springer, Boston, MA.

[15] 2019. *Deep-learning segmentation*. RaySearch Labratories AB.

[16] Ronneberger, O., Fischer, P., & Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation.

[17] Long, J., Shelhamer, E., & Darrell, T. 2014. Fully convolutional networks for semantic segmentation.

[18] Russell, S. J. & Norvig, P. 2010. *Artificial Intelligence: A Modern Approach, 3rd ed*. Pearson Education.

[19] Rich, E. 1983. *Artificial Intelligence*. McGraw-Hill.

[20] Naqa, I. E., Ruijiang, L., & Murphy, M. J. 2015. *Machine Learning in Radiation Oncology: Theory and Applications*. Springer International Publishing.

[21] Samuel, A. L. 1959. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210–229.

[22] Chollet, F. 2017. *Deep learning with Python*. Manning.

[23] Varoquaux, G. 2020. Scipy lecture notes, scikit-learn: machine learning in Python. URL: http://scipy-lectures.org/packages/scikit-learn/index.html.

[24] Raschka, S. & Mirjalili, V. 2017. *Python Machine Learning - Second Edition.*, volume 2nd ed. Packt Publishing. URL: http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1606531&site=ehost-live.

[25] LeCun, Y., Bengio, Y., & Hinton, G. 2015. Deep learning. *Nature*, 511, 436–444.

[26] Ciresan, D., Ueli, M., Masci, J., Gambardella, L. M., & Schmidhuber, J. 2013. Flexible, high performance convolutional neural networks for image classification. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2, 1237–1242.

[27] Goodfellow, I., Bengio, Y., & Courville, A. 2016. *Deep Learning*. MIT Press. URL: http://www.deeplearningbook.org.

[28] Grégory, D., Alexandre, E., Timothée, R., Julien, D., Jimmy, F., Caroline, N., Supiot Stéphane, L. T., & David, P. 2016. Comparison of automated atlas-based segmentation software for postoperative prostate cancer radiotherapy. *Frontiers in Oncology*, 6, 178.

[29] Lustberg, T., van Soest, J., Gooding, M., van der Stoep, J., van Elmpt, W., & Dekker, A. 2018. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiotherapy and Oncology*, 126(2), 312–317.

[30] Men, K., Dai, J., & Li, Y. 2017. Automatic segmentation of the clinical target volume and organs at risk in the planning ct for rectal cancer using deep dilated convolutional neural networks. *Medical Physics*, 44(12), 6377–6389.

[31] Tong, N., Gou, S., Yang, S., Ruan, D., & Sheng, K. 2018. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Medical Physics*, 45(10), 4558–4567.

[32] International Agency for Research on Cancer - World Health Organization. 2018. Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. URL: https://www.who.int/cancer/PRGlobocanFinal.pdf.

[33] Kreftregisteret. 2019. Cancer in norway. URL: https://www.kreftregisteret.no/Generelt/Rapporter/Cancer-in-Norway/cancer-in-norway-2018/.

[34] National Cancer Institute. 2019. Breast cancer. URL: https://www.cancer.gov/types/breast/patient/breast-treatment-pdq#_125.

[35] Kreftforeningen. 2019. Brystkreft - cancer mammae. URL: https://kreftforeningen.no/om-kreft/kreftformer/brystkreft/.

[36] Offersen, B., Boersma, L., Carine, K., Hol, S., Aznar, M., Biete, A., Kirova, Y., Pignol, J.-P., Remouchamps, V., Verhoeven, K., Weltens, C., Arenas, M., Gabryś, D., Kopek, N., Krause, M., Lundstedt, D., Marinko, T., Montero, A., Yarnold, J., & Poortmans, P. 01 2015. Estro consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. *Radiotherapy and Oncology*, 114.

[37] Helsedirektoratet. 2019. Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av pasienter med brystkreft. URL: https://www.helsedirektoratet.no/retningslinjer/brystkreft-handlingsprogram.

[38] Knuth, F. 2019. Automatic tumor delineation using mri and machine-learning. In *Proceedings of ESTRO 38*. European Society for Radiotherapy and Oncology.

[39] Clinicaltrials.gov identifier: Nct02541435. URL: https://clinicaltrials.gov/ct2/show/NCT02541435.

[40] Vaassen, F., Hazelaar, C., Vaniquia, A., Gooding, M., van der Heyden, B., Canters, R., & van Elmpt, W. 2020. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology*, 13, 1–6.

[41] Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302.

[42] Taha, A. A. & Hanbury, A. 2015. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*, 15, 29.

[43] Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. 1993. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9), 850–863.

[44] MathWorks. December 2019. Box plots. URL: https://se.mathworks.com/matlabcentral/answers/43629-citing-a-matlab-document.

[45] Vinod, S. K., Min, M., Jameson, M. G., & Holloway, L. C. 2016. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. *Journal of Medical Imaging and Radiation Oncology*, 60(3), 393–406.

[46] Gaede, S., Olsthoorn, J., Louie, A. V., Palma, D., Yu, E., Yaremko, B., Ahmad, B., Chen, J., Bzdusek, K., & Rodrigues, G. 2011. An evaluation of an automated 4d-ct contour propagation tool to define an internal gross tumour volume for lung cancer radiotherapy. *Radiotherapy and Oncology*, 101(2), 322 – 328.

[47] Conson, M., Cella, L., Pacelli, R., Comerci, M., Liuzzi, R., Salvatore, M., & Quarantelli, M. 2014. Automated delineation of brain structures in patients undergoing radiotherapy for primary brain tumors: From atlas to dose–volume histograms. *Radiotherapy and Oncology*, 112(3), 326 – 331.

[48] Dong, X., Lei, Y., Wang, T., Thomas, M., Tang, L., Curran, W. J., Liu, T., & Yang, X. 2019. Automatic multiorgan segmentation in thorax ct images using u-net-gan. *Medical Physics*, 46(5), 2157–2168.

[49] Mamani, G. E. H., Setio, A. A. A., van Ginneken, B., & Jacobs, C. 2017. Organ detection in thorax abdomen CT using multi-label convolutional neural networks. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, 287 – 292. SPIE.

[50] Trullo, R., Petitjean, C., Ruan, S., Dubray, B., Nie, D., & Shen, D. 2017. Segmentation of organs at risk in thoracic ct images using a sharpmask architecture and conditional random fields. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 1003–1006.

[51] Yang, J., Veeraraghavan, H., Armato III, S. G., Farahani, K., Kirby, J. S., Kalpathy-Kramer, J., van Elmpt, W., Dekker, A., Han, X., Feng, X., Aljabar, P., Oliveira, B., van der Heyden, B., Zamdborg, L., Lam, D., Gooding, M., & Sharp, G. C. 2018. Autosegmentation for thoracic radiation treatment planning: A grand challenge at aapm 2017. *Medical Physics*, 45(10), 4568–4581.

[52] Gooding, M. J., Smith, A. J., Tariq, M., Aljabar, P., Peressutti, D., van der Stoep, J., Reymen, B., Emans, D., Hattu, D., van Loon, J., de Rooy, M., Wanders, R., Peeters, S., Lustberg, T., van Soest, J., Dekker, A., & van Elmpt, W. 2018. Comparative evaluation of autocontouring in clinical practice: A practical method using the turing test. *Medical Physics*, 45(11), 5105–5115.

[53] Young, A. V., Wortham, A., Wernick, I., Evans, A., & Ennis, R. D. 2011. Atlas-based segmentation improves consistency and decreases time required for contouring postoperative endometrial cancer nodal volumes. *International Journal of Radiation Oncology, Biology, Physics*, 79(3), 943 – 947.

[54] van der Veen, J., Willems, S., Deschuymer, S., Crijns, W., Maes, F., & Nuyts, S. 2019. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiotherapy and Oncology, Volume*, 138, 68–74.

[55] Reed, V. K., Woodward, W. A., Zhang, L., Whitman, G. J., Buchholz, T. A., & Dong, L. 2018. Automatic segmentation of whole breast using atlas approach and deformable image registration. *Radiotherapy and Oncology, Volume*, 73(5), 1493–1500.

[56] Schreier, J., Attanasi, F., & Laaksonen, H. 2019. A full-image deep segmenter for ct images in breast cancer radiotherapy treatment. *Frontiers in Oncology*, 9, 667.

[57] Bauer, S., Nolte, L.-P., & Reyes, M. 2011. Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011*, Fichtinger, G., Martel, A., & Peters, T., eds, 354–361. Springer Berlin Heidelberg.

[58] Artan, Y., Haider, M. A., Langer, D. L., van der Kwast, T. H., Evans, A. J., Yang, Y., Wernick, M. N., Trachtenberg, J., & Yetik, I. S. 2010. Prostate cancer localization with multispectral mri using cost-sensitive support vector machines and conditional random fields. *IEEE Transactions on Image Processing*, 19(9), 2444–2455.

[59] Dong, M., Lu, X., Ma, Y., Guo, Y., Ma, Y., & Wang, K. 2015. An efficient approach for automated mass segmentation and classification in mammograms. *Journal of Digital Imaging*, 28, 613–625.

[60] Hanna, G., Hounsell, A., & O'Sullivan, J. 2010. Geometrical analysis of radiotherapy target volume delineation: a systematic review of reported comparison methods. *Clinical Oncology*, 22(7), 515 – 525.

[61] Norsk Forening for Medisinsk Fysikk (NFMF), Norsk Radiografforbund (NRF), Norsk onkologisk forening (NOF), KVIST. Inntegningscase fra norsk stråleterapimøte 2018 – tema brystkreft. Not published.

[62] Wong, J., Fong, A., McVicar, N., Smith, S., Giambattista, J., Wells, D., Lovedeep, G., Kolbeck, C., Giambattsita, J., & Alexander, A. 2019. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiation Oncology*, 105(1).

[63] Helseregisterloven (2001). lov om helseregistre og behandling av helseopplysninger (lov-2001-05-18-24). URL: https://lovdata.no/dokument/NL/lov/2014-06-20-43.

[64] Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. 2019. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. *Brainlesion*, 11383, 92–104.

[65] Jochems, A., Deist, T. M., van Soest, J., Eble, M., Bulens, P., Coucke, P., Dries, W., Lambin, P., & Dekker, A. 2016. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - a real life proof of concept. *Radiotherapy and Oncology*, 121(3), 459–467.

[66] Clark, K., Vendt, B., Smith, K., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., & Prior, F. 2013. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6), 1045–1057.

A Python script for extracting dose values in RayStation

This script is written in Python and was used in RayStation to extract different dose measures for a selection of ROIs in a treatment plan. The script creates an Excel spreadsheet with the dose values for each ROI for a chosen patient.

```
import clr, sys
from connect import*

clr.AddReference('Office')
clr.AddReference('Microsoft.Office.Interop.Excel')

import Microsoft.Office.Interop.Excel as interop_excel
import System.Array

excel = interop_excel.ApplicationClass(Visible=True)
workbook = excel.Workbooks.Add()
worksheet = workbook.Worksheets.Add()

# Create two-dimensional array
def create_array(m,n):
        dims = System.Array.CreateInstance(System.Int32,2)
        dims[0] = m
        dims[1] = n
        return System.Array.CreateInstance(System.Object, dims)

try:
        patient = get_current('Patient')
        plan = get_current('Plan')
except:
        print 'Patient␣and␣plan␣are␣not␣loaded.␣Exits␣script.'
        sys.exit()

# Set up header row
header_row = create_array(1, 4)
header_row[0,1] = 'Averge␣dose/V18␣Gy'
header_row[0,3] = 'V20␣Gy'

startcell = worksheet.Cells(1, 1)
header_range = worksheet.Range(startcell, startcell.Cells(header_row.GetLength(0),
    header_row.GetLength(1)))
header_range.Value = header_row

data_array = create_array(4, 5)
data_array[0,0] = patient.Name
data_array[0,1] = "Man"
data_array[0,2] = "AI"
data_array[0,3] = "Man"
data_array[0,4] = "AI"
```

```
data_array[1,0] = 'Heart'
data_array[2,0] = 'Left␣lung'
data_array[3,0] = 'Right␣lung'

rois = ["Heart", "Lung_left", "Lung_right"]
rois_AI = ["Heart_AI", "Lung_left_AI", "Lung_right_AI"]

summedDose = patient.Cases[0].TreatmentDelivery.FractionEvaluations[0].DoseOnExaminations
    [0].DoseEvaluations[0]

# Average dose to heart
data_array[1, 1] = summedDose.GetDoseStatistic(RoiName="Heart", DoseType = "Average")
data_array[1, 2] = summedDose.GetDoseStatistic(RoiName="Heart_AI", DoseType = "Average")

# V18 Gy and V20 Gy to left lung
data_array[2, 1] = summedDose.GetRelativeVolumeAtDoseValues(RoiName="Lung_left",
    DoseValues=[1800])[0]
data_array[2, 2] = summedDose.GetRelativeVolumeAtDoseValues(RoiName="Lung_left_AI",
    DoseValues=[1800])[0]
data_array[2, 3] = summedDose.GetRelativeVolumeAtDoseValues(RoiName="Lung_left",
    DoseValues=[2000])[0]
data_array[2, 4] = summedDose.GetRelativeVolumeAtDoseValues(RoiName="Lung_left_AI",
    DoseValues=[2000])[0]

# V18 Gy and V20 Gy to right lung
data_array[3, 1] = summedDose.GetRelativeVolumeAtDoseValues(RoiName="Lung_right",
    DoseValues=[1800])[0]
data_array[3, 2] = summedDose.GetRelativeVolumeAtDoseValues(RoiName="Lung_right_AI",
    DoseValues=[1800])[0]
data_array[3, 3] = summedDose.GetRelativeVolumeAtDoseValues(RoiName="Lung_right",
    DoseValues=[2000])[0]
data_array[3, 4] = summedDose.GetRelativeVolumeAtDoseValues(RoiName="Lung_right_AI",
    DoseValues=[2000])[0]

startcell = worksheet.Cells(2,1)
data_range = worksheet.Range(startcell, startcell.Cells(data_array.GetLength(0),
    data_array.GetLength(1)))
data_range.Value = data_array
worksheet.Columns.AutoFit()
```

B Python script for calculating quantitative metrics in RayStation

This script is written in Python and was used in RayStation for calculating overlap indices and spatial distance based metrics for two sets of ROIs that were to be compared. The script creates an Excel spreadsheet with the calculated DSC, HD, precision, sensitivity, specificity, and AVD for each ROI for a chosen selection of patients.

```
from connect import*
import clr, sys

clr.AddReference(’Office’)
clr.AddReference(’Microsoft.Office.Interop.Excel’)

import Microsoft.Office.Interop.Excel as interop_excel
import System.Array

# Create two-dimensional array
def create_array(m,n):
        dims = System.Array.CreateInstance(System.Int32,2)
        dims[0] = m
        dims[1] = n
        return System.Array.CreateInstance(System.Object, dims)

# Load patient database
patient_db = get_current(’PatientDB’)

# All relevant patients have ID 19061996
try:
    info = patient_db.QueryPatientInfo(Filter={’PatientID’:’19061996’})
except:
    print("Could␣not␣find␣patient␣info")

data_array = create_array(50, 50)

for i in range(len(info)):
    patient = patient_db.LoadPatient(PatientInfo=info[i])
    plan = patient.Cases[0].TreatmentPlans[0]
    structure_set = plan.GetStructureSet()

    Lung_left, Lung_right = ’Lung_left’, ’Lung_right’
    for r in structure_set.RoiGeometries:
        if r.OfRoi.Name == ’Lung_Right’ and r.HasContours():
            Lung_right = ’Lung_Right’
    for r in structure_set.RoiGeometries:
        if r.OfRoi.Name == ’Lung_Left’ and r.HasContours():
            Lung_left = ’Lung_Left’

    data_array[i, 0] = patient.Name
    # Extract evaluation metrics for heart, left lung and right lung
    # Metrics: 'DiceSimilarityCoefficient', 'Precision', 'Sensitivity', 'Specificity', '
        MeanDistanceToAgreement', 'MaxDistanceToAgreement'

    data_array[i, 1] = structure_set.ComparisonOfRoiGeometries(RoiA=Heart, RoiB='Heart_AI
        ', ComputeDistanceToAgreementMeasures=False)['DiceSimilarityCoefficient']
    data_array[i, 2] = structure_set.ComparisonOfRoiGeometries(RoiA=Lung_left, RoiB='
        Lung_left_AI', ComputeDistanceToAgreementMeasures=False)['
        DiceSimilarityCoefficient']
    data_array[i, 3] = structure_set.ComparisonOfRoiGeometries(RoiA=Lung_right, RoiB='
        Lung_right_AI', ComputeDistanceToAgreementMeasures=False)['
        DiceSimilarityCoefficient']

file_path = None
close_excel = True

try:
    # Open Excel with new worksheet
    excel = interop_excel.ApplicationClass(Visible=True)
    workbook = excel.Workbooks.Add(interop_excel.XlWBATemplate.xlWBATWorksheet)
    worksheet = workbook.Worksheets[1]

    # Set up header row
    header_row = create_array(1,5)
    header_row[0, 0] = 'Patient'
    header_row[0, 1] = 'Heart'
    header_row[0, 2] = 'Left␣lung'
    header_row[0, 3] = 'Right␣lung'

    # Add header row to work sheet
    startcell = worksheet.Cells(1, 1)
    header_range = worksheet.Range(startcell, startcell.Cells(header_row.GetLength(0),
        header_row.GetLength(1)))
    header_range.Value = header_row

    # Add ROI data array to work sheet
    startcell = worksheet.Cells(2,1)
    data_range = worksheet.Range(startcell, startcell.Cells(data_array.GetLength(0),
        data_array.GetLength(1)))
    data_range.Value = data_array

    # Auto-fit the width of all columns
    worksheet.Columns.AutoFit()

finally:
    # The following is needed for the excel process to die when user closes worksheet
    if file_path != None and close_excel:
        excel.Quit()
    System.Runtime.InteropServices.Marshal.FinalReleaseComObject(worksheet)
    System.Runtime.InteropServices.Marshal.FinalReleaseComObject(workbook)
    System.Runtime.InteropServices.Marshal.FinalReleaseComObject(excel)
    seriesCollection = None
    chart = None
    worksheet = None
    workbook = None
    excel = None
    System.GC.WaitForPendingFinalizers()
    System.GC.Collect()
```

C Normality test

Figure 42 to 44 show some of the results of the Shapiro-Wilk tests, which was used to assess whether the differences in dose values, described in chapter 3.1.3, follow a normal distribution. The samples in Figure 42 and 44 can be assumed to follow a normal distribution (p > 0,05), while the sample in Figure 43 cannot be assumed to follow a normal distribution (p < 0,05).

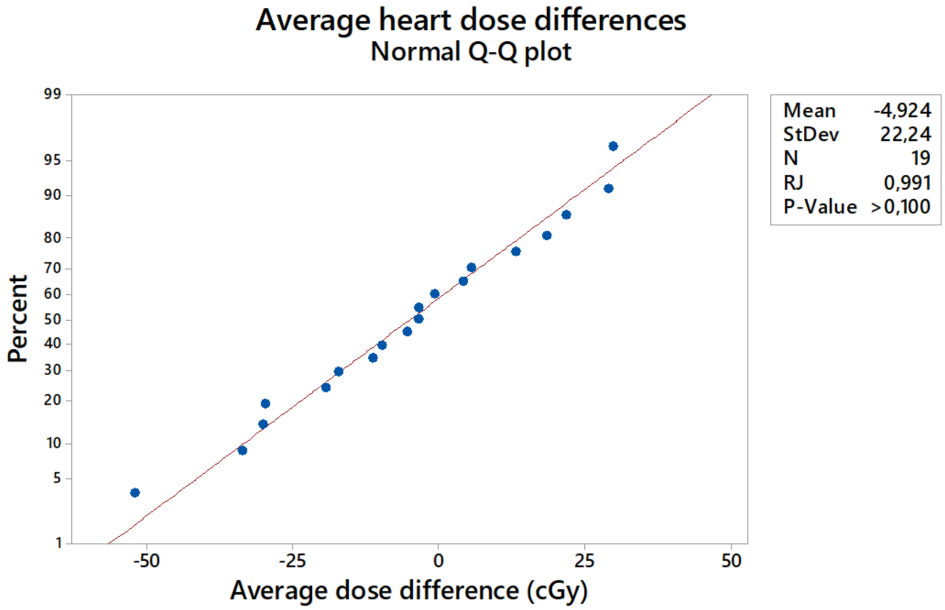


Figure 42: Normal Q-Q plots for the average heart dose differences between clinical and AI segmentations obtained with the DL thorax model.

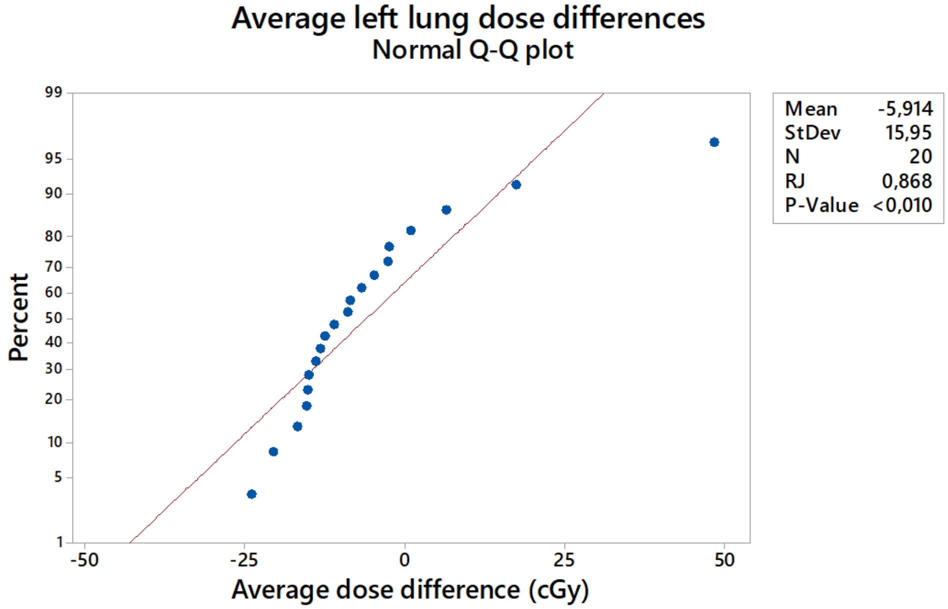


Figure 43: Normal Q-Q plots for the average left lung dose differences between clinical and AI segmentations obtained with the DL thorax model.

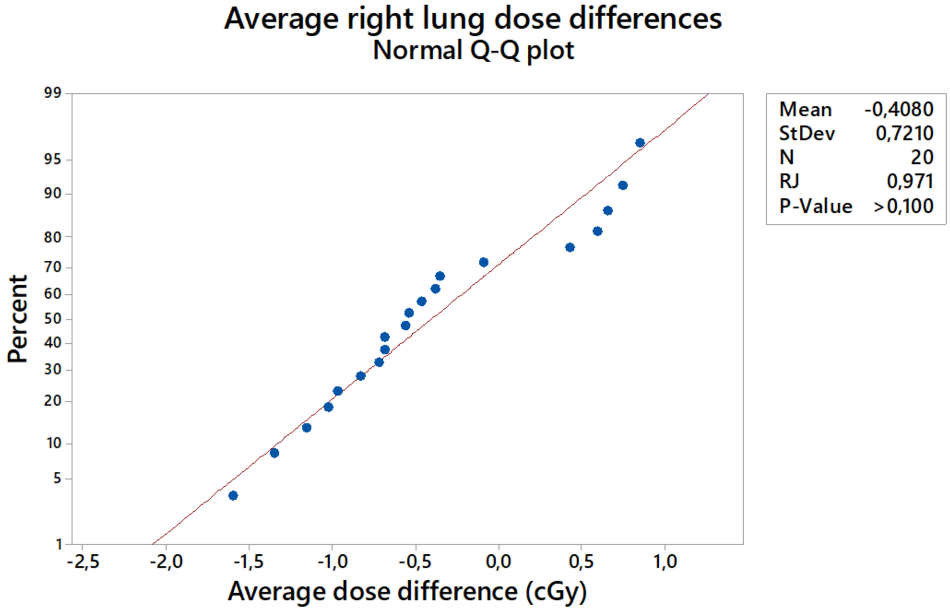


Figure 44: Normal Q-Q plots for the average right lung dose differences between clinical and AI segmentations obtained with the DL thorax model.

NTNU
Norwegian University of
Science and Technology

ST. OLAVS HOSPITAL
TRONDHEIM UNIVERSITY HOSPITAL