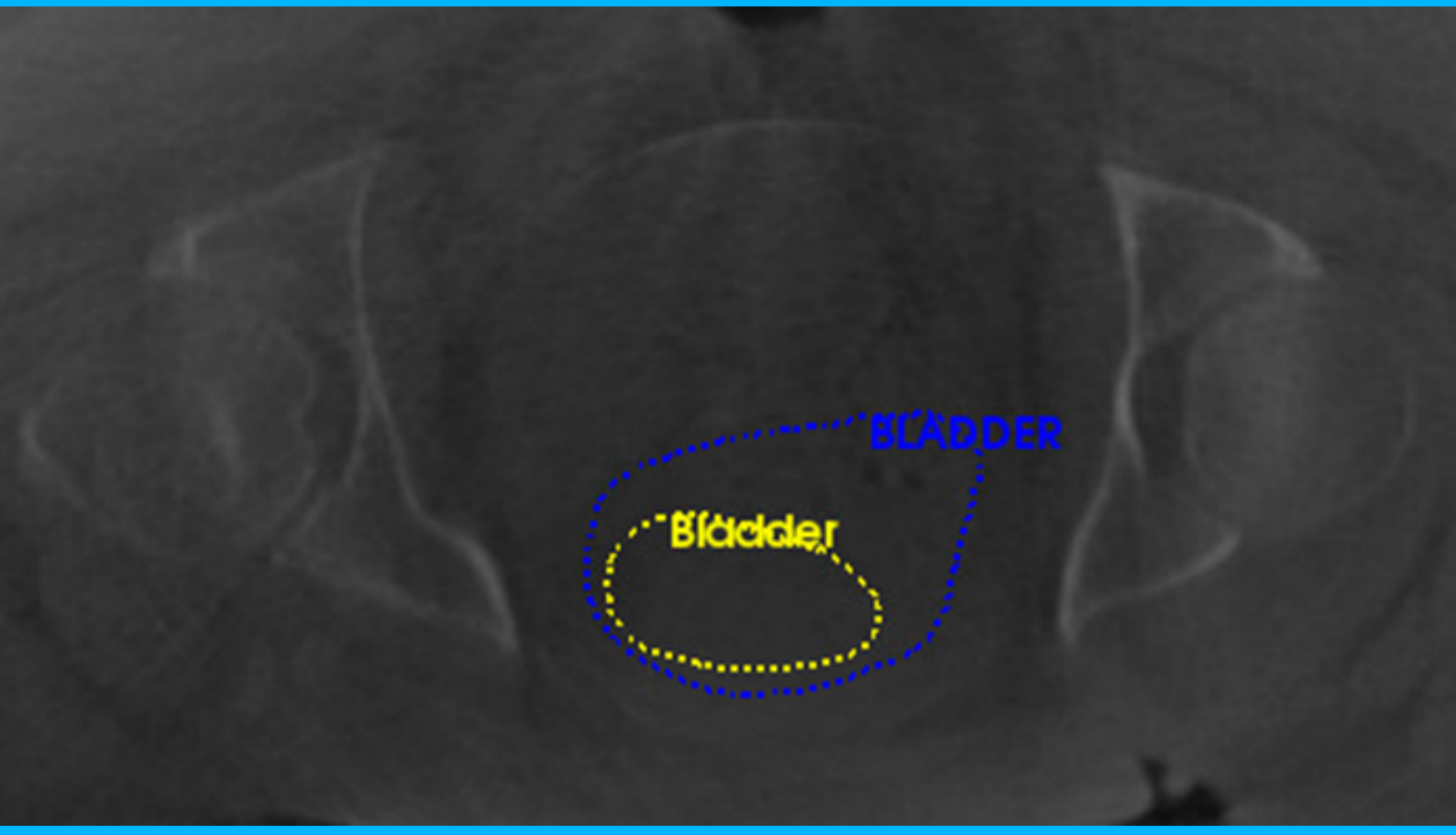
Automatic Contour Quality Assurance on CBCT scans for Locally Advanced Cervical Cancer Patients

A comparison study using Machine Learning

María Teresa Ruiz Alba



TU Delft

Erasmus
ERASMUS
UNIVERSITEIT
ROTTERDAM

Automatic Contour Quality Assurance on CBCT scans for Locally Advanced Cervical Cancer Patients

A comparison study using Machine Learning

by

María Teresa Ruiz Alba

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday January 18, 2022 at 11:00 AM.

Student number:	5132134	
Thesis committee:	dr. ir. D. R. Schaart,	TU Delft
	dr. ir. J. Schiphof-Godart,	Erasmus MC
	prof. dr. M.S. Hoogeman,	TU Delft and Erasmus MC
	M.Sc. D. Reijtenbagh,	Erasmus MC

TU Delft

Acknowledgements

First of all, thank you Jérémy for letting me dive into this very interesting project. I have learnt a lot during the process. Thank you for your weekly supervision, feedback, and the valuable knowledge I take home. I would also like to thank Dennis for accepting to supervise my thesis and for transmitting your enthusiasm in our meetings. Thank you also to Mischa for giving me very valuable feedback, and for that enriching discussion that made me think out of the box.

I am especially grateful to Dominique and her constant support during the whole thesis. Thank you for all the feedback and advice you have given me, for always being available, and for our conversations not related to the thesis, they have helped me a lot in every aspect.

Finally, thank you to my family for always supporting me no matter what, to my friends in Delft for being my family away from home, and to my friends in Spain for always being there throughout the years.

María Teresa Ruiz Alba
Delft, January 2022

Abstract

Background and purpose: One of the main challenges in external beam radiotherapy treatment of locally advanced cervical cancer patients is dealing with bladder and rectum filling. Organ filling causes inter-fraction motion of the uterus, requiring large treatment planning volumes, or a plan library. Current assessment of tumor position is mainly done by visual inspection of a Cone Beam Computed Tomography (CBCT) scan. Eventually, this can lead to inter- and intra-observer variability when choosing the best treatment plan from the plan library based on bladder filling. The incoming introduction of auto-contouring tools to obtain automatically-generated (AG) contours of the bladder and the rectum on CBCT scans, allows the easier identification of these organs at risk and consequently, faster localization of the tumor region. However, to rely on these AG contours in the decision of plan selection, it is necessary to know if they have been reliably segmented. The goal of this project is to develop a strategy based on quantitative image features, to evaluate the quality of the AG contours to know if they are suitable for plan selection assessment.
Materials & Methods: 140 LACC patients from Erasmus MC were included. For each patient, bladder and rectum contours were obtained from each of the CBCT scans done throughout the treatment (five fractions (CBCT scans) per patient). These contours were automatically-generated using a deep learning-based auto-segmentation algorithm. Gold-standard contours were manually delineated in some CBCT scans, but the rest of the automatically-generated contours did not have the corresponding ground-truth contour, hence they were labeled with a score between 1 (bad quality) and 5 (good quality). For consistency, gold-standard contours were included in the dataset with the class label 5. The contours were relabeled to have a binary classification problem, and those with label 3 were removed. Each contour volume was divided into three subregions: core region, inner and outer shell. This contour data was used for a comparison study between two supervised machine learning (ML) methodologies: Random forest (RF) networks and Logistic Regression (LR). For both strategies, feature extraction and selection were implemented. In RF methodology, a prior step of dimensionality reduction using principal component analysis (PCA) was performed. In LR, univariate feature selection followed by a multivariate logistic regression analysis was done. Before implementing the classifiers, the dataset was split into a training set and a test set. The ML models were trained using the training set, and they were tested on new unseen data. Predictions on the test data were obtained and used for evaluation of the models performance using evaluation metrics: accuracy, sensitivity, specificity, confusion matrix, ROC curve and AUC.
Results: The RF classifier performed on the bladder test data with an AUC value of 0.87, while for the LR model, the value obtained was 0.77. The trained RF model identified the accurate and inaccurate bladder contours with a sensitivity of 94% and a specificity of 54%. The trained LR model resulted in a sensitivity of 91% and a specificity of 42%. In the case of the rectum, the RF classifier performance is indicated with the AUC value of 0.89, while the LR model obtained a value of 0.84. In the case of sensitivity and specificity, the RF model got 96% and 38%, and the LR classifier 95% and 38%, respectively.
Conclusion: Random forest classifiers give the best results in terms of performance and classification skills for the OARs considered, especially for the bladder. It has been demonstrated that quantitative image features, paired with the corresponding contour class label, can be used for deriving statistical relationships from the data. This allows the identification of contouring errors and classifying the contours based on their quality. With the increasing automation of different steps in the radiotherapy treatment workflows, the automatic contour QA tool developed would be a key step in the process to ensure a faster, more feasible and consistent plan selection. The tool could act as a support tool for radiotherapy technicians when choosing the plan from the plan library that best fits the daily anatomy of the patient.

Contents

List of Tables

List of Figures

1 Introduction

Cervical cancer is one of the most common malignancies affecting women around the world. In 2020, the disease was diagnosed in more than half a million women worldwide, leading to more than 300,000 deaths [1]. In the Netherlands, almost 800 women are diagnosed with cervical cancer each year, and more than 200 women die from it [2].

Cervical cancer originates from the cervix, the lower part of the uterus, which makes it part of the female reproductive system. When a patient is diagnosed with an invasive carcinoma in the cervix, staging of the cancer is performed according to the classification made by the International Federation of Gynecology and Obstetrics (FIGO) [3]. The table describing the staging of cervical carcinoma can be found in the references [3, 4]. The FIGO staging ranks the degree of cervical cancer spread, being the stage I when the tumor is strictly confined to the cervix with a maximum size of 4 cm. Stage IV represents the worst degree of cervical cancer spread: stage IVA is when the cancer is spread to adjacent pelvic organs or regional metastases, while stage IVB occurs when the carcinoma has spread to distant organs, which may end up affecting the bones, lungs or liver [4]. A patient is diagnosed with locally advanced cervical cancer (LACC) when an invasive carcinoma bigger than 4 cm is detected and starts spreading to areas beyond the uterus up to stage IVA [4]. The patient cohort used in this project only includes LACC patients.

1.1. Cervical cancer treatment

The treatment of cervical cancer with curative intentions involves surgery, radiotherapy, or a combination of chemotherapy and radiotherapy [5]. Surgery is the treatment of choice when the tumor is confined to the cervix. However, for LACC patients the tumor has spread beyond the cervix and has started to invade the adjacent pelvic organs, therefore surgery is no longer an option. External beam radiotherapy (EBRT) with concurrent chemotherapy and image-guided adaptive brachytherapy are the current choice for LACC treatment [6]. In the worst stage of the carcinoma (stage IVB) patients are offered chemoradiotherapy as a palliative treatment for the metastatic disease to distant organs [7].

1.1.1. External beam radiation therapy for LACC patients

In radiation therapy, cancer cells are killed by delivering high energy x-rays to the target areas. EBRT and brachytherapy are two radiation therapy techniques used during the treatment of LACC patients. In EBRT the radiation is delivered from the outside of the patient towards the inside, directed to the tumor region. Concurrent chemoradiation is usually part of the treatment as well, and brachytherapy may be also given to the patient following EBRT and chemoradiation [8]. Brachytherapy, also known as internal radiotherapy, consists on locally delivering a high dose to the tumor by temporarily introducing a radioactive source inside the patient's body in or near the tumor [9].

In general, LACC patients' treatment consists of 25 fractions for delivering a total dose of 45 Gy to the tumor and surrounding regions. Therefore, the patient is irradiated with 1.8 Gy per fraction. The total treatment time is between 5 to 7 weeks [10]. The target volumes are the cervix, uterus, parametrium and pelvic lymph nodes, while the organs at risk (OAR) are the bowel, bladder, rectum, sigmoid, bowel,

and femoral heads. The kidneys and spinal cord are included in the list only if para-aortic irradiation occurs [11].

Intensity Modulated Radiation Therapy (IMRT) and Volumetric Modulated Arc Therapy (VMAT) are two modern EBRT techniques. They provide an improvement regarding organ at risk (OAR) sparing, and achieving more conformal dose distributions when compared to the conventional 3D conformal radiotherapy (3DCRT) [12]. VMAT compared to IMRT provides a further improvement of OAR sparing and a shorter treatment delivery time [10, 13].

1.1.2. Challenges of cervical cancer radiotherapy treatment

The potential benefits of IMRT and VMAT over 3DCRT for cervical cancer treatment are limited by the geometrical day-to-day variations in the pelvic area, affecting the precision of the treatment delivery [10].

These daily geometrical variations are a challenge for ensuring conformal radiation delivery to the target volume. Especially bladder and rectum filling have a large impact on the daily cervix-uterus shape and position [14].

Earlier research gathered in the systematic review published by Jadon et al. [10] showed that within the target volume for radiation delivery, the uterine motion is larger than the cervical motion. Moreover, cervical and vaginal motion are mainly caused by rectal filling (see Figure 1), while bladder filling has more influence on uterine motion (see Figure 3). Especially the motion in anterior-posterior (AP) direction of the uterus' fundus has been found to be largely affected by bladder filling [14]. When measuring the displacement in the upward direction, Buchali et al. found that the uterus can move up to 15 mm and the cervix up to 6 mm due to bladder filling [15]. These anatomical variations may compromise the dose distribution initially planned for the target, with the consequent detrimental effect on target coverage. A straightforward solution for this problem would be to increase the irradiated target volume. However, even though the priority in radiation therapy is ensuring tumor coverage, it would lead to dramatically reduce healthy tissue sparing, with the consequent increased toxicity for the patient.

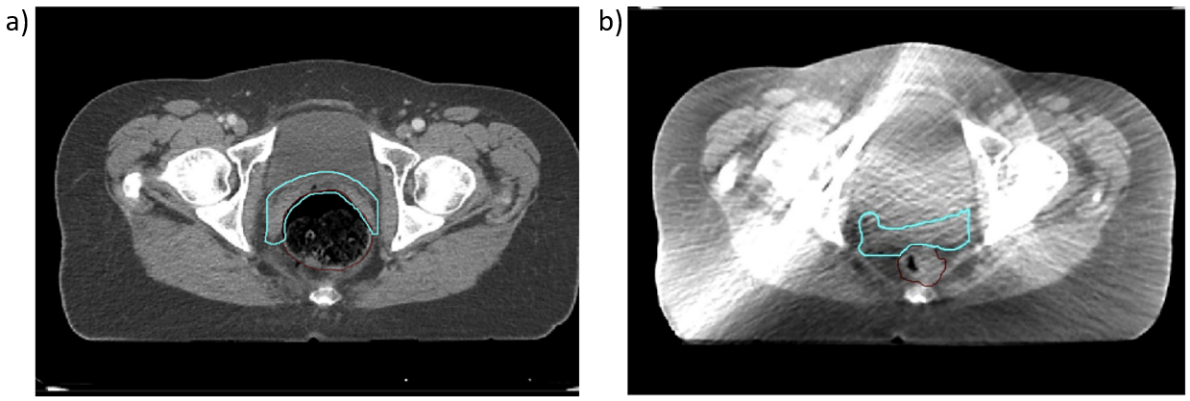


Figure 1: Axial view of the vaginal change in position due to rectal filling. (a) Planning CT showing the vaginal CTV (blue) affected by the distended rectum (brown). (b) CBCT scan obtained during the treatment of the same patient showing how the vaginal position has changed to more posterior due to a smaller rectum [10]. The difference in image quality is noticeable: planning CT (a) shows a better contrast and resolution; CBCT scan (b) is affected by artifacts and the poor image contrast makes it difficult to distinguish the target volumes and OARs.

To deal with the unavoidable organ motion during LACC treatment, Ahmad et al. [16] stated the proof of principle for a methodology consisting on a model that could be used for predicting the position of the cervix-uterus during treatment. This was done by creating a pre-treatment motion model using correlation between bladder-filling variations and cervix-uterus displacements.

This pre-treatment cervix-uterus motion model was used by Bondar et al. [17] to create an improved workflow for LACC radiation therapy. They proposed a system of plan libraries consisting of treatment plans created from different Internal Target Volumes (ITVs). These ITVs were obtained from delineating different cervix-uterus positions each linked to different bladder volume range.

This system of plan libraries was implemented in Erasmus Medical Center (Erasmus MC) in 2011, introducing in clinical practice an online adaptive Plan-of-the-Day (PotD) protocol. This PotD protocol is the current approach implemented for the EBRT treatment part of LACC patients, and it consists on generating a patient-specific plan library that covers the full motion range of the cervix-uterus as a function of bladder filling.

Treatment planning and delivery are the two mains steps that need to be followed to decide the EBRT treatment for each LACC patient. In the next sections, the workflow followed in Erasmus MC will be explained further, specially looking into the PotD protocol.

1.1.3. Treatment planning: PotD protocol

The main steps followed during treatment planning are shown in Figure 2. Before plan library generation, the radiation oncologist places polymer-based markers in the patient. This allows the easier identification of the cervix in a Computed Tomography (CT) scan, hence finding more easily the tumor. Then, two planning CT scans are obtained for each patient: a full and an empty bladder CT scan, which are aligned on the bony anatomy by a radiotherapy technician (RTT) to a previously acquired diagnostic Magnetic Resonance (MR) scan [7]. This MR scan is useful for easier identification of structures, especially the tumor since magnetic resonance imaging (MRI) provides better contrast for soft tissue.

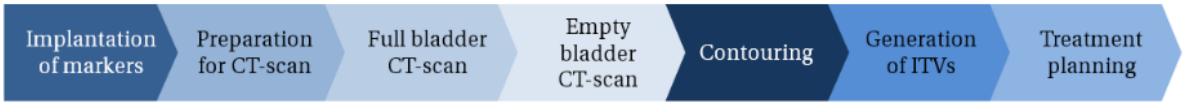


Figure 2: Treatment planning process for cervical cancer patients implementing the PotD protocol. *CT = Computed Tomography, ITV = Internal Target Volume*. Schematic reprinted from Anouk Corbeau's master's thesis [7].

Contouring of the target structures and OARs as well as treatment planning are done on the planning CT scans, which are used to generate model-predicted Internal Target Volumes (mpITVs) [18]. These ITVs encompass the motion of the uterus due to bladder filling, and they are the basis for creating the library of treatment plans. The delineation methodology and the dose constraints for these OARs are specified in the EMBRACE-protocol [11].

The range of bladder filling-induced motion of the cervix-uterus can vary a lot among patients. Therefore, the number of treatment plans generated for each patient's plan library depends on the displacement measured at the tip of the uterus. If the displacement is more than 2.5 cm, the patient is considered a large mover, otherwise, it is a non-mover (or small mover), see Figure 3 [18]. For large movers, the plan library consists of two VMAT treatment plans with small margins, one from empty-to-mid-full bladder, and the other one from mid-full-to-full bladder. An additional backup plan with bigger margins completes the plan library but is based on the empty-to-full bladder motion. In the case of small movers, only one VMAT plan is generated in addition to the backup plan [18]. Therefore, the plan library for movers consists of 3 treatment plans, while for non-movers it consists of 2 treatment plans.

Sometimes the patient presents an anatomy-of-the-day that doesn't fit any of the small margin plans in the plan library. In this case, the motion-robust backup plan, with more generous margins, is chosen for treatment delivery to ensure tumor coverage. The backup plan is also chosen when the image quality of the Cone Beam CT (CBCT) scan is very poor (the use of the CBCT scan is introduced in section 1.1.4). However, this implies higher dose delivered to the OARs, hence increasing toxicity [19]. In Erasmus MC, when the backup plan is chosen more than three times during the whole treatment, the radiation oncologist decides whether a new treatment plan library should be generated [7].

Once the patient-specific plan libraries are generated from the mpITVs, a treatment planning system is used to design a radiation dose distribution that is optimized and deliverable for the anatomy of each patient, and their specific dose requirements. The radiation oncologist is in charge of reviewing and approving the treatment plans [7].

1.1.4. Treatment delivery

During each treatment fraction, before delivering the radiation dose, an in-room CBCT scan of the patient is acquired. Rigid registration between the CBCT image and the full bladder planning CT scan is done using the pelvic bones as reference. This allows the repositioning of the patient in case of

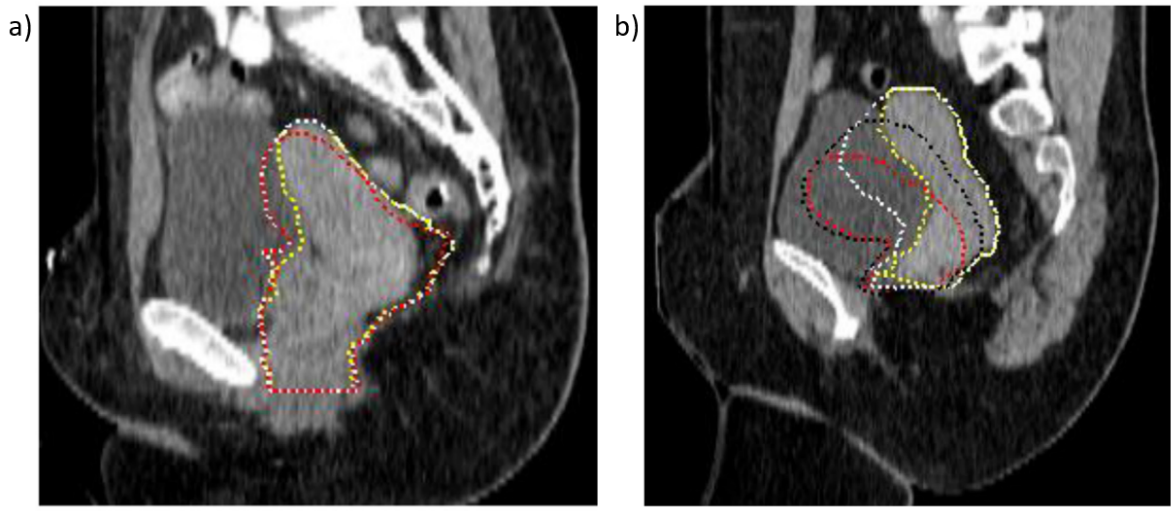


Figure 3: Sagittal view of full bladder CT scans from two LACC patients. The cervix-uterus position is delineated in the full bladder CT scan (yellow), and empty bladder CT scan (red). (a) Non-mover patient showing a small cervix-uterus motion. The white delineation is the mpITV containing the whole range of motion of the cervix-uterus due to bladder filling. (b) Patient with large cervix-uterus motion (mover). The empty-to-half-full mpITV is shown in black, and the half-full-to-full mpITV in white [18].

detecting a large rotational deviation of the pelvic and vertebral bones between the CBCT and planning CT scan. If translation is detected, it can be corrected by simply moving the patient table. Then, the CBCT scan is used to select the plan from the patient-specific plan library that best fits the anatomy-of-the-day, based on the bladder filling. Plan selection mainly depends on the experience of the RTTs and on visually checking how the anatomy shown in the CBCT scan matches any of the plans in the library [7].

When delivering the radiation dose to the patient in the form of high-energy photon beams, the medical linear accelerator (LINAC) is the most commonly used device for EBRT. As previously mentioned in section 1.1.1, IMRT and VMAT are advanced treatment delivery techniques for photon therapy, and both are considered the standard radiotherapy techniques used for cervical cancer treatment [11].

1.2. Research Question

During treatment planning, the OARs and the tumor target are manually delineated by an experienced clinician. The most common approach for contour quality assurance (QA) is by human examination, which can lead to missing detectable contouring errors due to avoidable factors like fatigue, or increasing workload [20]. With daily treatment, there is no time for manual delineation, and visual inspection of the contours is a time-consuming process that requires human expertise, especially for sub-optimal CBCT image quality, and that can lead to inter and intra-observer variability within and across radiotherapy centers [21].

However, plan library expansion has been investigated to improve healthy tissue sparing while maintaining target coverage, potentially reducing treatment-related side effects. It has been shown that for patients with large bladder-induced motion of the cervix-uterus, expanding the plan library from 2 to 3 or 4 treatment plans is beneficial [22]. However, creating more treatment plans per patient increases the workload of the RTTs and the decision-making time for plan selection.

The implementation of online auto-contouring of the OARs on the daily CBCT scans would act as a supporting tool for automatic plan library selection based on the anatomy-of-the-day. By finding the treatment plan that best matches the automatically delineated contours, plan selection would be faster and would not rely on human expertise. However, the quality of these automatic delineations may be affected by the low-quality images of the CBCT scans. Therefore, a contour error detection tool for quality assurance becomes necessary to know if these auto-generated contours are good enough for treatment plan selection assessment.

The purpose of this study is to find a strategy for quantitatively evaluating the quality of automatically

delineated contours on CBCT scans of LACC patients, to assess their eligibility for aiding plan selection. This methodology should be based on quantifiable aspects of the images, rather than solely relying on the qualitative evaluation by RTTs. Since cervix-uterus motion is mainly influenced by bladder and rectum filling, only the delineations of these two OARs are considered in the project. Therefore, features describing characteristics of the bladder and the rectum have been studied, and statistical relationships between these descriptors and the quality of their corresponding contours have been inferred. These statistical relationships learned from previous data, would allow us to check whether a new unseen automatically-generated contour is good enough or not for plan selection.

1.3. Thesis structure

This thesis dives into the problem explained in this introduction with the purpose of finding an answer to our research question. To do so, the following chapters form this report: Chapter 2 provides a review of the literature giving a deeper insight into the research gap for understanding the strategies used in this study. Chapter 3 provides a description of the dataset used and the two methodologies implemented, i.e. dimensionality reduction followed by a random forest classifier, and feature selection through univariate and multivariate logistic regression analysis; Chapter 4 shows all the results obtained with these methodologies for the delineations of the two OARs considered, i.e. bladder and rectum, showing that the random forest classifier performed better than the logistic regression model for both OARs overall. In Chapter 5 a discussion of the results and approaches implemented is provided as well as the limitations of this project and future work following this research, leading to a final conclusion in Chapter 6.

2

Literature Review

This chapter is an adaptation of the report written for the literature review with title "*Methodologies for quality assurance of automatically-generated contours for locally advanced cervical cancer patients*" performed before starting this thesis project. The purpose of this literature study is to review how the process for contour error detection is currently done for cervical cancer patients. Then, the existing methodologies for automating the contour QA procedure during radiotherapy treatments are investigated.

2.1. Literature search strategy

The literature study was conducted following the systematic review procedure, with PubMed and Scopus as the databases chosen. There were no constraints set on the publication date due to the limited number of articles in the field of study. The last review of the literature was done in October 2021. The search strategy designed to find the most relevant publications was focused on finding articles dealing with automatic contour QA for cervical cancer patients. However, as almost no studies were found, it was decided not to limit the search to cervical cancer patients and explore the methodologies developed for other types of cancer instead. Therefore, the search query used was: *"automat*" AND "contour*" AND (("quality" AND ("assurance" OR "assessment" OR "validation")) OR "error detection") AND "radiotherapy"*.

Only the publications implementing geometric or feature-based methodologies were considered. Studies evaluating the clinical effect of delineation uncertainty by analyzing its dosimetric impact have not been included in this literature review.

In PubMed, 57 publications resulted from the search, against the 97 found in Scopus. When comparing the studies obtained from both databases, 50 duplicates were found and removed. A total number of 104 unique publications were left. Figure 4 shows the procedure followed for selecting the appropriate literature in detail.

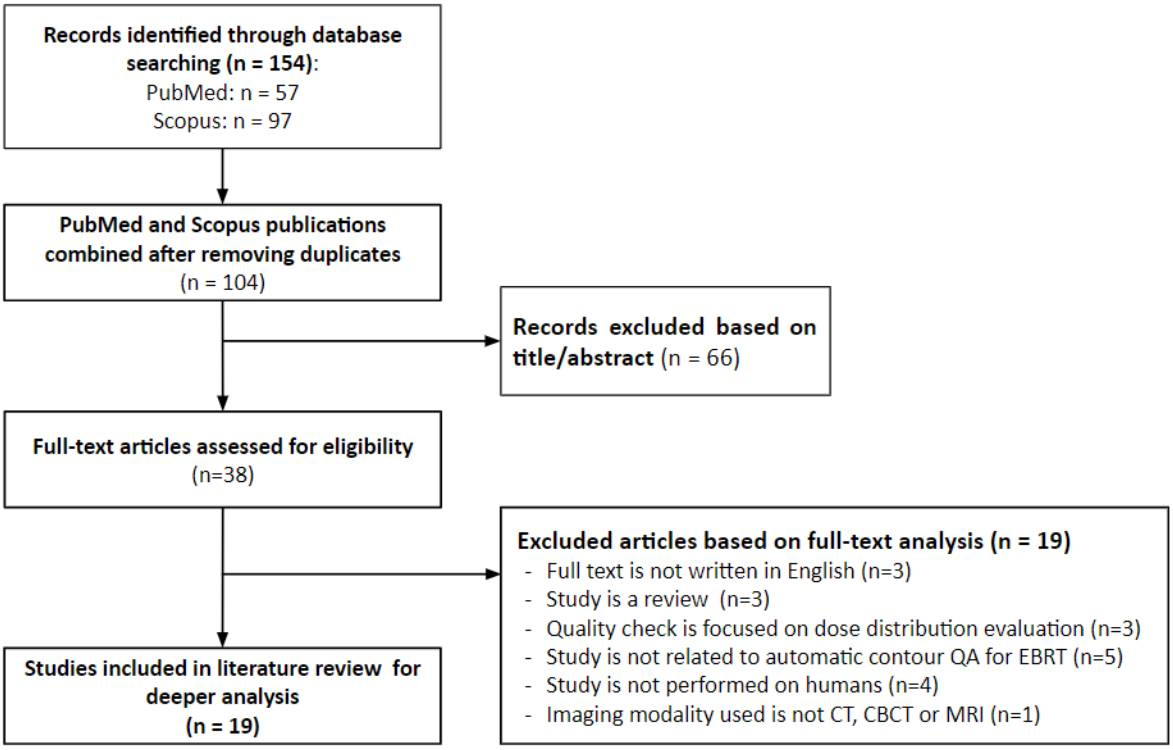


Figure 4: Flowchart showing the process for reviewing the literature.

2.2. Methodologies for quality assurance of automatically-generated contours

Several methodologies have been encountered in the elaboration of this literature review regarding automatic contour QA. First, a review on studies implementing commonly used quantitative metrics as their strategy for contour error detection is presented. Then, publications involving Artificial Intelligence (AI) are explored.

As the current contour QA consists of a visual check by RTTs, the methodologies explained in the sections 2.2.1 and 2.2.2 are not the clinical standard practice but mostly methodologies in the research stage. The investigation of these new strategies originated from the commonly noticed need for standardizing the contour QA process while making it more robust and consistent.

2.2.1. Commissioning of auto-segmentation algorithms

The process of commissioning auto-segmentation algorithms used to obtain automatically-generated contours generally involves the implementation of a set of commonly used quantitative metrics, which are mainly overlap- and distance-based metrics [23]. In this context, these similarity metrics need a gold-standard (GS) contour that acts as a benchmark for comparison with the automatically-generated (AG) contour. The former is usually manually segmented by an experienced clinician, while the latter is the one whose quality needs to be assessed.

Overlap-based metrics

Overlap metrics measure the volume or surface overlap between the reference or GS contour and the AG contours. Dice Similarity Coefficient (DSC) is the most used metric for validating medical volume segmentations. It ranges from 0, meaning no overlap among any of the voxels in the contour volume, to 1, which indicates complete overlap [24]. Considering that we have two sets of contours (A and B), the formula is as follows:

$$DSC = \frac{2|A \cap B|}{|A| + |B|} = \frac{2TP}{2TP + FP + FN} \quad (1)$$

where TP denotes the True Positives, i.e. the AG contours correctly labelled as accurate; FP represents the False Positives, meaning the AG contours incorrectly classified as accurate or with good quality; FN indicates the False Negatives, which are the AG contours that are correctly delineated but are labelled as inaccurate.

Overlap metrics are fast to calculate and straight forward to implement. However, they only focus on the comparison of the contour volumes without specifically considering the border of the contoured organ, and they are also less sensitive to larger volumes [25].

Distance-based metrics

Distance metrics measure the distance between the AG contour and the GS contour, and they are used as additional descriptors of contour characteristics to assess their segmentation quality. Hausdorff Distance (HD) is the most commonly used metric of this type, and is defined as follows:

$$HD(A,B) = \max\left(h(A,B),\ h(B,A)\right) \tag{2}$$

$$h(A,B) = \max_{b\in B}\left(\min_{a\in A}||a-b||\right) \tag{3}$$

where h(A, B) is the largest distance from a point in contour A to the nearest point in contour B [24].

An advantage of distance-based metrics is that they take into account the spatial location of false positives and false negatives. However, they are complicated to interpret for small-volume contours [26].

Several papers have been published where both types of geometric indices are used to evaluate the performance of their auto-segmentation algorithms (mainly, deep learning-based) for multiple cancer sites. Some examples are head and neck cancer [25, 27, 28, 29], breast cancer [30, 31], rectal cancer [32], prostate cancer [33] and cervical cancer [34]. The structures automatically delineated in these studies include both, the OARs and in some cases the tumor region.

However, before using these metrics for auto-segmentation validation, it is important to consider their limitations. Metrics that are sensitive to point positions like the HD are more suitable than volume-based metrics, such as the DSC. Moreover, using volumetric and overlap similarities as the criteria for contour assessment is not the best practice when the quality of the segmentation is low [26]. CBCT images have poor contrast and low quality, making it difficult for the auto-segmentation algorithm to generate high-quality contours. Therefore, in the case of only implementing these traditional evaluation metrics, the best approach would be to select a combination of overlap and distance metrics, since it has been shown that they are often not correlated [35].

These publications use the mentioned metrics as a validation step to measure the accuracy of their segmentation algorithm. However, they are not enough to ensure that the AG contours are good or not, since they merely depend on comparing the geometry of the delineations, while additional aspects that are hidden in the image should be explored to generate better descriptors of contour quality. Especially for cervical cancer patients, the tumor can change its shape and position due to bladder and rectum filling, which explains why these metrics are not enough as a unique QA methodology. More advanced strategies should be explored to study different aspects of the contours, giving more insights into the characteristics that define them. In the next section, the methodologies found in the literature are explained.

2.2.2. AI-based automatic contour QA

In section 2.2.1, the studies presented require a set of GS contours to benchmark the AG contours. They are mainly focused on giving a qualitative evaluation of the performance of the auto-segmentation algorithms implemented. However, assessing contour quality using these traditional evaluation metrics is not always possible due to the lack of ground-truth contours. Moreover, these metrics are not enough as a methodology for assessing contour quality since their ability to generalize to other cancer sites subjected to more deformations, as is the case of cervical cancer, is more limited.

In recent years, applications of AI in medical physics have grown exponentially. In particular, the use of Machine Learning (ML), as a branch of AI, and Deep Learning (DL), as a subfield of ML, have been studied for medicine and oncology over a wide range of applications from image segmentation, to knowledge-based planning, quality assurance and radiomics feature extraction [36].

More specifically, ML consists on building statistical models and using algorithms to perform certain tasks by analyzing and drawing inferences from patterns in the input data, without being explicitly programmed to perform those tasks [36]. There are three types of learning inside ML: supervised learning, unsupervised learning, and reinforcement learning [37]. Only supervised learning will be explained further due to its relevance to the topic researched in this literature review.

Supervised learning uses a labeled dataset to teach models the desired output (label) depending on the input variable. These algorithms are expected to learn the relationship between the input data and the labels to be able to guess the corresponding output for new unseen data [36].

DL algorithms are built around the idea of how the brain works and its structure. These algorithms are called artificial neural networks, and in contrast to classical ML algorithms, they are capable of extracting features from the raw data without the need of getting into the feature engineering process of feature extraction and selection [36].

These techniques provide advanced automation of the contour QA process making it more robust and consistent. The publications that have been explored in this section concern ML- and DL-based strategies for automatic contour QA, and are further analyzed below.

Machine Learning-based methods

Several studies have been published showing how ML and in particular, supervised ML, can be used for automatic contour QA to identify contouring errors in the target structures.

In general, these studies have a set of AG contours that are either labelled with a score or with a pass/fail describing their conformity to the target structure. Then, statistical relationships between these labels and contour quality are inferred from ML models and are learned and generalized from the given dataset so that they can be later used for contour classification of new unseen data. This means that the approaches presented in this section are based on probabilities of the contours belonging to a class. In the light of supervised ML, the classification problem relies on previously labeled data to assign a certain class (accurate vs. inaccurate) to the contour [38].

In 2013, McIntosh et al. identified errors in contours from multiple Regions of Interest (ROIs) in the thorax, including tumor targets and OARs, by extracting geometric and voxel intensity-based features from CT images. These features were analyzed with a model based on Conditional Random Forests. The methodology implemented solves a classification problem framed as one of conditional probabilities, and consists on getting the probability of a contour from belonging to a certain class, given its set of features, using sets of decision trees as the learning algorithm [39]. The ultimate goal is to make sure that the ROIs segmented in the radiotherapy treatment plan have been correctly labelled, ensuring that the quality of the plan is good enough for radiation delivery.

Chen et al. presented in 2015 an automatic contour QA strategy based on geometric attribute distribution models which were used for OAR contour error detection in head and neck cancer patients, using CT images. The geometric attributes considered were the centroid, volume, and shape of the structures evaluated. Moreover, the spatial relationship between adjacent OARs, and the anatomical contour similarity among different patients were also included. The results obtained for the average sensitivity and specificity were promising, achieving a low false detection rate, hence showing that the implemented strategy is feasible for contour error detection [40]. The strategy developed had the goal of alleviating the part of the physicists' workload which consists in visually reviewing OAR contours before treatment planning, ultimately improving the radiation therapy workflow.

Altman et al. (2015) proposed a solution for contour QA designed to be part of an online adaptive radiotherapy (OL-ART) workflow, and to validate manually or automatically-generated contours, reducing treatment time and variability. For this purpose, a knowledge-based strategy was implemented using historical data from head and neck cancer patients consisting of CT images and their corresponding delineations. A total of 9 different contours per CT image were obtained: brain, brainstem, eyes, optic chiasm, optic nerves, and parotid glands. Metrics like shape, size, and relative position (centroid-to-centroid distance) were calculated for all the contours in the study and used as descriptors. Clinically relevant information like sex, age, and weight was also included in the features data set. Population statistics for each metric were obtained from the knowledge-based data and were used to set the decision criteria to classify the input patient data. The decision criteria were not tailored for each contour class, but were instead applied to all the structures using a window of +/-1.96σ, where σ is the standard deviation of each metric. This window was used as "passing" criteria: if the metrics computed for each delineation did not pass the threshold previously determined for each of them, the contour would

be classified as inaccurate. The knowledge-based contour QA tool reported a sensitivity of 95% and specificity of 81% [41].

Zhang et al. (2016) implemented an automatic contour QA tool for OL-ART of prostate cancer patients, with the goal of speeding up the process of error detection in AG contours [42]. The input data consists of daily CT images acquired with a CT-on-rails, and their corresponding correct prostate, bladder and rectum contours. This data was used for building a geometry model expressing the shape of each structure. Principal component analysis (PCA) and Procrustes analysis were used during model construction with the purpose of ensuring an efficient and feasible segmentation of the target area and the OARs:

- PCA is a dimensionality reduction technique that aims at reducing the feature space by finding the main eigenvectors (principal components) that best explain the variance in the given data set, removing correlation between variables. PCA is mainly a rotation of the data to a new reference system based on the principal components while preserving the variation between data points [43].
- Procrustes analysis is a statistical method used to analyze the distribution of a given set of multidimensional shapes and transform them so that maximal superimposition is achieved. Shapes are aligned and differences between structures are minimized so that only the real shape variations are measured. In Procrustes analysis, data is translated, scaled and rotated to a common coordinate system [43].

For contour error detection, an adaptive decision tree was implemented to analyze the test shape compared to the previously obtained model shape. Promising results were achieved, suggesting that using the automatic contour QA tool for online adaptive radiotherapy could become an option. However, the authors noted that improvements on the speed and robustness of the QA tool should be made, especially when looking at the algorithm and chosen features of the organs [42].

Hui et al. (2018) developed a QA tool based on comparing volumetric features from the OARs in CT images, and a statistical reference obtained from historical priors containing features from lung cancer patients [20]. All features were computed in three-dimensional space in contrast to the approach presented by Altman et al. (2015), where the computation of the features was mainly two-dimensional. The general workflow implemented includes a statistical inference test, which is set up using the parametric distribution of each volumetric feature. The QA strategy consists of identifying the outlier features that do not correspond to the previously set outlier criteria, which is based on the parametric distributions in the initial OAR reference. Errors in delineations are reported if an abnormal parametric distribution of a feature is detected. If the evaluated OAR contour is classified as normal, then it is included in the OAR reference set.

The user's improvement on error detection sensitivity using the QA tool was measured, and showed that the proposed methodology mostly had a significant impact on improving major error detection. To define what a major/minor error was, three experienced radiation oncology residents reviewed all the contours and classified the identified errors. The possible impact of the detected contour errors on treatment planning and dosimetry was evaluated by each of the radiation oncology residents, and based on their own opinion and considerations, the detected errors were classified as being either major or minor through majority voting.

The sensitivity for minor error detection remains relatively low only increasing from 61% to 68% using the QA tool. However, in the case of major errors, the improvement from 78% to 87% shows that the QA tool can prove quite valuable in accurate detection of major contouring errors [20].

More recently, Terparia et al. (2020) proposed an approach comparing different ML models and using six Conformity Indices (CIs): DSC, Jaccard Conformity Index (JCI), van't Riet Index (VRI), Geographical Miss Index (GMI), Discordance Index (DI), and HD. The agreement between the gold-standard and the evaluated contours was visually checked, and the latter were labelled accordingly by an experienced RTT with either "pass" or "fail" [44].

CI values were calculated for each pair of gold-standard and evaluated contours and were later matched to their corresponding pass/fail label. Different ML models were trained and tested on the data, giving as input the CI values for each contour with their corresponding mapped pass/label score; 70% of the data was used for training and 30% for testing. The ML models considered in this study were Decision Tree, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbour (KNN)

and Ensemble, which consists of the combination of different ML models to obtain a better one with increased performance. The dataset consisted of 393 contours corresponding to tumor volumes and OARs: liver Gross Tumor Volume (GTV), node GTV, spine GTV, liver, oesophagus, stomach, and heart. The ML models were trained using as features the previously mentioned CIs for each contour, and as the labels their corresponding pass/fail score qualitatively determined. Three different approaches were followed to train the ML algorithms:

- All 393 contours were considered a single group.
- Tumor volumes and OARs contours were separated into two different groups.
- Each structure was considered an individual group.

The results indicated that better values for predictive accuracy, sensitivity and specificity are obtained with the third approach, i.e., having one model per structure. Another important conclusion from this publication is that assessing contour conformity using CIs is not the best approach, since they have different values between structures. There are no criteria specifying the CI values that would correspond to high and low conformity for a particular structure. Hence ML models are used to infer the statistical relationships between the quantitative data and the labels. However, the problem is that the metrics used are not the best descriptors of contour conformity [44].

Moreover, JCI, DSC and VRI are mathematically comparable, which suggests that these CIs evaluate very similar aspects of each pair of contours leading to correlated results. The authors provide scatter plots showing the correlation between pairs of these features. Therefore, for further improvement of this methodology, two of these three correlated metrics should be discarded as predictors [44].

Up to now, most of the machine learning-based methodologies presented rely on geometrical or location-based features. The downside of these descriptors is that they are constrained to the shape and position of the contour. Zhang et al. (2019) proposed an innovative methodology for contour error detection on structures subjected to shape and positional changes during the radiotherapy treatment [45]. Through this contour QA strategy, the goal was to make more manageable the inter-fraction deformations of the OARs studied, making the implementation of online adaptive replanning easier. Pancreatic cancer patients were used in this study, such that for each of them the pancreas head and duodenum were manually delineated by an experienced physicist and double-checked by a radiation oncologist, creating the set of accurate (gold-standard) contours. The delineations were obtained from CT images acquired before radiation delivery of each fraction, using an in-room CT-on-rails.

Deformable-image-registration-based contour propagation was used for creating a set of inaccurate contours. From each organ mask contour, three different subregions were obtained: core, inner shell, and outer shell. From each subregion, 38 texture features were extracted and dimensionality reduction with PCA was performed on the set of accurate contours to find the principal features explaining 95% of the variance. Based on some texture constraints determined by the three subregions, a decision tree model with three levels was built. The dataset was divided into the training and test sets, consisting of two-thirds and one-third of the data respectively. The decision tree model was trained on the training set and tested on the test set with the purpose of classifying contours as either accurate or inaccurate. Only the contours meeting the passing criteria for every level of the decision tree were classified as accurate. The decision criteria were set based on the feature distributions of the training set learnt by the decision tree. If a contour is classified as inaccurate, it is reviewed and classified again [45].

The method developed in this publication for automatic contour quality assessment was proven to be effective and feasible for organs that experience significant deformations between fractions during the radiotherapy treatment. Moreover, the authors suggested that the model would perform even better in other cancer sites experiencing less variations in their shape and position, or with better image contrast between structures, as for example head and neck cancer patients.

This is the first methodology presented that could set the basis for standardizing the contour quality assessment procedure. The created workflow does not solely focus on geometrical or location-based features, but also considers image texture features, which creates a more flexible strategy for cancer sites subjected to tumor and organ motion during the radiotherapy treatment.

Deep Learning-based methods

In recent years, DL has been introduced in radiotherapy through the development of automatic segmentation algorithms for tumor targets and OARs, helping to save time and reduce inter- and intra-observer variability. However, even if the best DL model is used for automatic delineation, slice by slice contour inspection or modification by physicians is still needed to ensure contour quality before radiation delivery, which is very time-consuming [46].

Automatic contour QA is very important to ensure a feasible and efficient radiotherapy treatment. DL-based methodologies for contour QA are presented in this section. Even though the number of publications for DL-based automatic segmentation strategies is extensive, the studies specifically related to the use of DL for automatic contour QA are very limited.

Moreover, the publications found and explained in this section present automatic contour QA methodologies suitable for fully adaptive radiotherapy. However, the research gap investigated in this literature review concerns the QA process of the contours after treatment planning generation, i.e., for plan selection.

Chen et al. presented in 2020 the first DL-based methodology for automatic contour QA. The group of patients used in this study were 680 early-stage breast cancer patients, 340 with left-sided breast cancer, and the remaining 340 with right-sided breast cancer. The ground-truth contours consist of manual Clinical Target Volumes (CTV) delineated on CT images by experts. The proposed strategy consists of a convolutional neural network (CNN) model which performs the auto-segmentation of the contours, followed by a QA network based on ResNet-101 [46]. These are the main steps proposed as the workflow:

- Perform automatic contour segmentation of two-dimensional (2D) CT images to obtain the corresponding 2D probability maps (*p*), in which each pixel value represents its probability of belonging to the contour being segmented. Since this study was focused on providing an automatic contour QA methodology instead of a new method for auto-segmentation, an existing CNN with demonstrated high performance was used.

- Use the segmentation probabilities to calculate the 2D uncertainty maps (*u*), which will be used for prediction of segmentation quality, as shown in Equation 4, where (*i, j*) denotes each pixel and the uncertainty of each pixel is denoted by *u(i, j)*:

$$u(i,j) = \begin{cases} p(i,j), & 0 \leq p(i,j) \leq 0.5 \\ 1 - p(i,j), & 0.5 < p(i,j) < 1 \\ 0, & p(i,j) = 1 \end{cases} \tag{4}$$

  The pixels that are close to the decision boundary of the segmentation have higher uncertainty values. When the probability of a pixel belonging to the segmented contour is 1, its uncertainty is 0. This shows how closely related the quality of the automatic segmentation is to the uncertainty map [46].

- Get predictions on segmentation quality using a classification model with inputs: 2D CT images, probability maps and uncertainty maps. The metric used to identify segmentation quality was the DSC and based on this, three different quality levels were defined in order to classify the contours. The range of DSC values defining each quality level were determined based on the authors experience and on a previous review of the literature:

  – Good quality (label 0): DSC value is in the range [0.95, 1].

  – Medium quality (label 1): DSC value is in the range [0.8, 0.95].

  – Bad quality (label 2): DSC value is in the range [0, 0.8].

  Two different outputs were tested for this QA network: one output predicts directly the DSC value for each slice, and the other output predicts the quality label. ResNet-101, the classification network used, consists of 101 convolutional layers that extract low, middle and high-level visual features. The last step of the network is a softmax layer, which classifies the predicted segmentation quality values into three categories (good, medium, bad), according to their predicted DSC value and the range in which it falls [46].

- Quantitatively assess the QA model performance regarding prediction of the quality levels using metrics like receiving operator characteristic curve (ROC), Area under the curve (AUC), F score, and balanced accuracy (BA), which quantifies the ability of a system to avoid false classification. For evaluating the performance of the model on predicting the DSC values, the mean absolute error (MAE) was the quantitative metric used.

- Revise the automatic segmentation. This is done by a physician.

A downside of this study that limits its applicability to other cancer sites is that this workflow was tested on the CTV of breast cancer patients, which shows very good contrast with surrounding tissues. Promising results were obtained, however this is a big constraint when considering how this workflow would work on low-quality images with poor contrast, as it happens with the CBCT scans of LACC patients. The authors point out that if the presented auto-segmentation model is tested on other tumor sites or OARs, it would show poor performance or even fail [46].

Moreover, using the DSC value as the index for segmentation quality is not the best approach since it is susceptible to varying absolute contour volume. This could have led to contour mislabeling because a small volume delineation will have a small DSC value, which does not always mean that the contour is inaccurate.

Men et al. (2020) proposed another DL-based automatic contour QA strategy, but this time for assessing multi-center OAR contouring of lung cancer patients [47]. The OARs evaluated in this study were the heart, esophagus, spinal cord, and lungs. Deep active learning was implemented for automatic segmentation following these steps:

- 110 cases from clinical trials were divided into three groups: candidate, validation and test sets.

- A gold-standard atlas with 36 cases was used for training a CNN segmentation model. The CNN took as input the CT images to output the corresponding segmentation probability maps for each OAR. A deficiency of this study was the training set size, since the authors were aware that it was not representative of the whole population.

- To deal with the limited training set, the CNN segmentation model was fine-tuned by adding images from the candidate set to the training set. To maximize the learning of the CNN model the images and corresponding contours selected were the ones with the highest uncertainty parameters, i.e. the images closest to the decision boundary of the model. The selection of the images was done based on their *"representativeness"*, which was computed by estimating the uncertainty and accuracy and combining both quantities into one parameter. Using a similar idea as Chen et al.(2020), Men et al. (2020) estimated the uncertainty value for each image from the pixel probabilities as shown in Equation 5:

$$U_n = \frac{1}{m}\sum_{1}^{m}(1 - \max(p_i, 1 - p_i)) \tag{5}$$

  where U_n represents the uncertainty value of the n-th image, *m* indicates the pixel index in the n-th image, and p_i is the probability of the i-th pixel of belonging to the OAR to be segmented [47].

  Combining the uncertainty (U_n) and the segmentation accuracy, which was computed using the DSC and the HD, a representativeness parameter (R_n) was obtained (Equation 6):

$$R_n = \frac{U_n \times DSC_n}{HD_n} \tag{6}$$

  where DSC_n and HD_n represent the DSC and HD value of the n-th image.

- After fine-tuning the segmentation model, contour accuracy was assessed with the validation set not only using the DSC metric, but also the HD, as was suggested by Chen et al. (2020) to improve the evaluation of the segmentation performance. DSC and HD were also used to establish the QA criteria.

For this purpose, the mean and the standard deviation (σ) of both evaluation metrics for each OAR were computed using the accurate contours in the validation set. Thresholds for DSC and HD were defined as the pass criteria as shown below:

$$DSC_{test} > mean_{DSC} - 1.96\sigma_{DSC} \tag{7}$$

$$HD_{test} < mean_{HD} + 1.96\sigma_{HD} \tag{8}$$

- The test set was used and the fine-tuned CNN model and the decision criteria (Equations 7 and 8) were applied to it to identify the inaccurate contours. Taking into account both metrics for contour classification makes this methodology more robust than the one proposed by Chen et al. (2020). Small contours usually have a small DSC value, and would thus have higher chances of being misclassified as inaccurate contours. However, small contours usually have better HD values. Therefore, to avoid these classification errors, the contours would only have to meet one of the set criterion to be classified as ”correct”, and become acceptable for treatment planning [47].

- The performance of the automatic contour QA model was quantitatively evaluated with the following metrics: sensitivity, specificity, ROC curve and BA.

The results showed that 95% of the contours were identified correctly without any further manual checking being necessary. Moreover, the QA strategy developed was able to detect slices with missing errors [47].

The main limitation is that the decision criteria used in this study were not investigated specifically for each OAR. Therefore, for some of the organs the decision criteria were not well defined and some contours that passed the thresholds may still need additional checking by physicians [47].

The publications using DL methodologies to implement an automatic contour QA strategy are limited and recent, mostly from 2020. Further research in this field must be carried out, especially regarding contour QA for cervical cancer patients, and on CBCT scans. No studies were found for LACC patients implementing DL for this purpose, nor performing automatic contour QA on CBCT scans.

2.3. Discussion

Multiple methodologies have been explored for automatic contour QA, which shows that a consensus on a standardized methodology is absent. In particular, studies using delineations on CBCT scans have not been found. Moreover, the research concerning LACC patients is very limited. Hence studies showing a methodology for contour QA for other cancer sites were considered. As shown in section 2.2.1, it is common practice to use overlap and distance metrics for evaluation and commissioning of auto-segmentation algorithms. However, they rely on having another set of GS contours, which sometimes is not possible since they are usually manually delineated. Moreover, they are not enough to ensure contour conformity, especially for structures that can deform significantly during the radiotherapy treatment.

Despite the popularity of these metrics for the assessment of auto-segmented contours, Cha et al. (2021) observed in their results that these geometric indices have a weak relationship with the quality scores previously given by physicians [33]. It was reported that the metrics used showed a limited ability to determine which contours had clinically significant errors, which is probably related to the fact that the prostate cancer patient cohort used is also affected by OAR displacement. This study concludes that efforts should be directed towards creating more consensus among clinicians and developing methodologies for improving contour QA. This gives additional evidence and motivation to find alternative strategies for contour error detection.

Moreover, studies that tackle the problem of contour QA for cancer patients that have tumors in mainly static areas, like head and neck, usually use as a tool for contour error detection overlap or distance metrics. However, for the case of LACC patients, the tumor may have moved considerably from one fraction to another. These displacements in the cervix-uterus position can lead to big differences in shape and position between the AG contours and the GS contours. Therefore, QA based on contour overlap with the gold standard defined at the beginning of the treatment (planning CT), or metrics based on distance (HD) are not reliable enough, since the position and even geometry of the contour may have changed.

This highlights the need for finding more generalizable parameters or features on which we can base the contour error detection strategy. Contours are boundaries in an image defining the limit between the inner and outer part of a structure, in our case, OARs like bladder and rectum. These edges or contours exist because of the sharp changes in pixel intensities. Hence, everything relies on what the gray values in the CBCT scan can tell us from the underlying structures. This is why studying the radiomics can help us extract the maximum potential out of medical images, since there is a lot of underlying information given by the pixel values, their organization, and the structures they form.

Therefore, features based on the distribution, position, geometry, and gray levels of these pixels can be extracted. Making decisions on contour quality based on this versatile information given by radiomics is a more robust and reliable technique than solely using geometric indices.

One of the big challenges of cervical cancer treatment is the indistinctive soft tissue boundaries, which are even harder to identify when the quality of the image is low with poor image contrast, as is the case of CBCT scans. Because studies using MRI scans benefit from the high soft tissue contrast of this imaging modality, their implemented methodology is not very applicable to the research question that this literature review investigates. If the image contrast is very good, the chances of having a good auto-segmentation are higher. Hence geometric and location-based features might be enough to compare the contours. Moreover, if the group of cancer patients considered is of the type of static cancers, like head and neck, less variations in location and shape of the target structures are expected between treatment fractions. Therefore the use of overlap and distance metrics to check for contouring errors is also common among these studies.

Regarding the use of DL for automatic contour QA explained in section 2.2.2, the described methodologies show how by giving the raw input images, the neural networks can learn the features that are useful for contour classification according to the classes defined. However, a drawback of using DL for this purpose is that we don't know on which features the classification is based on.

Moreover, DL models typically need much more training data to obtain feasible predictions than the more traditional ML models. All this data needs to be labeled, and in many cases, to deal with the heavy workload, data is labeled by non-experts or by automatic systems. The work by Men et al. heavily relied on having a gold standard atlas publicly available, which might not be the case for other cancer sites and OARs [47]. This means that a standard database for the target area would need to be manually prepared, which is very time-consuming.

Both DL-based methodologies explained rely on the calculation of an uncertainty parameter to evaluate the quality of the contours. However, as Chen et al. (2020) reflected in their study [46], it may be that the uncertainty is not directly related to the performance of the segmentation model. The uncertainty value expresses how confident the segmentation model is about each pixel belonging or not to the analyzed structure, but this may not be correlated with how accurate the prediction is. Instead, if the segmentation model has systematic errors in the same area, the resultant network confidence may be more affected by intra- and inter-observer variability of the contours in the dataset, as well as a limited training dataset that is not representative of the whole population [46].

As pointed out at the beginning of the DL subsection in 2.2.2, the explained studies develop a QA strategy for online adaptive radiotherapy. This allows for modifications of the contours each time the decision criteria are not met. The procedure is then repeated in a loop until the contour is considered accurate. The problem with DL-based contour QA pipelines is that the evaluation metric they rely on to assess contour quality is mainly DSC. However, if we apply this strategy to plan selection, the goal of automating this contour QA would not be met.

Plan selection needs an automatic tool that supports the clinician and provides a double-check on contour quality so that the RTT can make a faster and reliable decision on which treatment plan matches better the anatomy-of-the-day. Contours in the treatment plans of the plan library are already defined, and hence do not need to be modified to better fit the anatomy-of-the-day. Therefore, a strategy with features that allows us to efficiently determine contour quality is needed, which requires robust and non-correlated descriptors.

Therefore, as the study by Hui et al. (2018) emphasized [20], automatic contour QA tools should be seen as a reinforcement and assistance tool for contour error detection, not as the end point of the QA workflow. Medical physicists should still be the ones deciding about the validity of a contour, but having a supporting tool like the ones discussed in this review, would be of great help to reduce the workload, potentially improving the safety and quality of the radiation therapy treatment.

2.4. Conclusion

With the introduction of online auto-contouring in the clinical workflow, an unbiased and feasible protocol for contour quality assessment of the AG contours is needed. This would allow the process to rely less on human expertise, which can lead to inter- and intra-observer variability, and reduce treatment time.

Moreover, from all the reviewed studies, it can be observed that a current standard in radiotherapy for quantitative assessment of contour quality is lacking. To have a standardized methodology, the complexity of structure movement for certain types of cancer should also be taken into account. This suggests that to create a robust method, not only are geometric and volumetric features needed, but also image texture-based features. In addition to this, a model that can derive statistical relationships between these features and the quality labels would be the method of choice.

AI-based methodologies provide a more robust solution for contour QA. Moreover, interpretability by a medical physicist of the QA strategies used is important, since they are the bridge between these automatic QA tools and their clinical implementation. ML is easier to interpret than DL, and needs less volume of data to be labeled, which is beneficial for cancer sites that do not have an atlas of images and contours available. Therefore, for LACC patients, supervised ML is the technique that would provide a more robust and reliable contour QA strategy while keeping accessible the interpretability of the results for the medical physicists.

Finally, quantitative evaluation of the AG contours should be seen as a supporting tool on contour quality check, not as a sole decision-making tool. Clinicians should still review the output of these algorithms, however having a tool for automatic contour error detection already helps reducing the workload and the decision-making time for simpler cases by detecting gross-contouring errors, leaving more time for clinicians to perform a more in-depth evaluation of the most complex cases.

3

Materials & Methods

In this chapter, the methodologies investigated to answer the research question are explained. In section 3.1, the dataset used is described. In the following sections, the strategy used in this project is detailed: section 3.2 explains the steps followed for preparing the data for the rest of the workflow; section 3.3 describes how feature extraction was performed, and the steps that pre-processing of the feature data entailed; in section 3.4 the two supervised Machine Learning (ML) algorithms implemented are explained. Finally, Section 3.5 describes the metrics used for evaluating the performance of each ML model, with the final aim of comparing their outcomes and choosing the best methodology to solve the research question. Figure 5 shows a summary of the whole methodology implemented in the project.

Matterhorn, which is a software development platform at Erasmus MC, Python 2.7.12, and Matlab R2017a have been used for extracting the data and implementing the scripts used in this project.

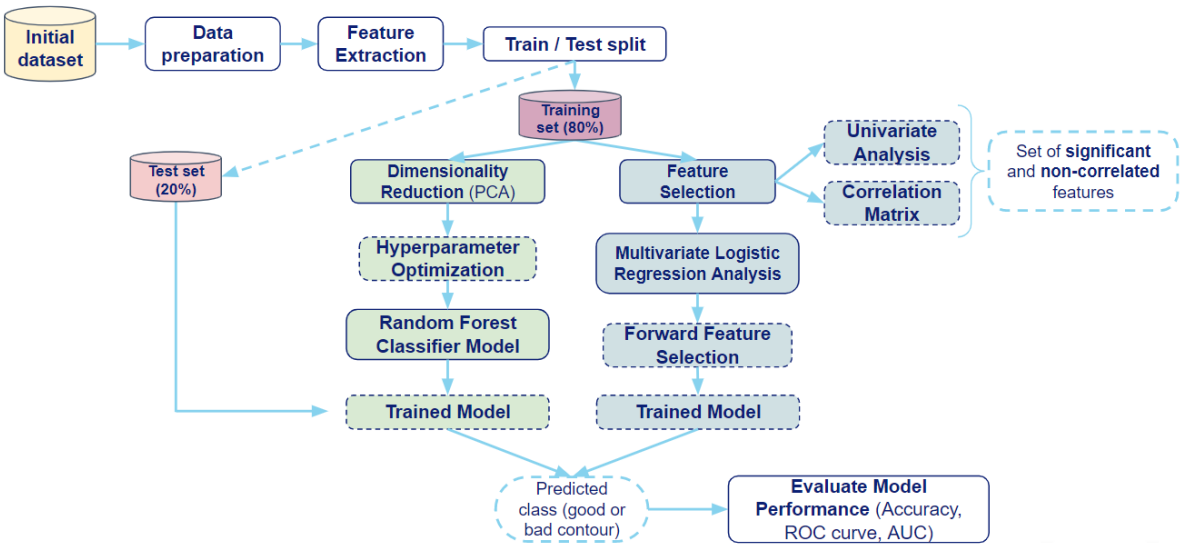


Figure 5: Diagram showing the general workflow implemented in this project.

3.1. Image and contour datasets

Previously anonymized data from 140 LACC patients was provided by Erasmus MC. This data consists of 5 CBCT scans per patient, corresponding to each of the 5 different fractions in the radiotherapy treatment, having a total of 710 CBCT scans. Due to errors found in the images, 8 CBCT scans were removed from the dataset, having 702 CBCT scans left for processing. These errors were the following:

- The automatically-generated contours were missing in the dataset due to poor quality of the image, hence the CBCT scans without delineations were deleted.
- The Field of View (FoV) is too small.
- The labels for some contours were missing, hence the corresponding CBCT scans were removed from the dataset.

3.1.1. Automatically-generated contours

Each of these CBCT scans contains one structure set, which includes the delineation of the bladder, the rectum, and the external (or skin) contour. These contours were automatically generated by a DL-based auto-segmentation algorithm. This project is focused on finding a methodology for assessing the quality of bladder and rectum contours. Therefore, the final dataset of AG contours does not include the skin delineations.

A properly trained Ph.D. student performed a visual evaluation of the quality of the AG contours and assigned a score from 1 to 5, grading their suitability for plan selection. The chosen scoring scale goes as follows: scores 1 and 2 correspond to very bad and bad contour quality, respectively; contours that were neither good enough nor too bad were assigned label 3; finally, labels 4 and 5 were given to delineations with high or very high quality, respectively.

3.1.2. Gold-standard contours

This project revolves around creating an algorithm or workflow that allows us to check whether an AG contour is good enough or not for plan selection. For this purpose, it is necessary to know what an accurate contour for the bladder and the rectum looks like and what their characteristics are. Therefore, the GS contour dataset needs to be created. The previously mentioned Ph.D. student manually delineated the contours for the bladder and the rectum for 136 CBCT scans. The remaining 566 CBCT scans were not delineated, hence, their structure sets do not have a gold-standard. The reference on contour quality that we have for the rest of the AG contours is the scoring label that was given to them.

The GS contours are called like that because they are the ground truth or reference of how a high-quality delineation should be. Therefore, to match the labeling system used in the project, they were included in the dataset paired with the scoring label 5.

Figure 6 shows an example CBCT scan for the bladder with the AG contour and the corresponding GS contour from a patient included in the described dataset. The slices selected show the big difference between both delineations in a particular region of the organ volume. The auto-contouring tool used for creating the AG contour was probably affected by the poor image contrast of the CBCT scan.

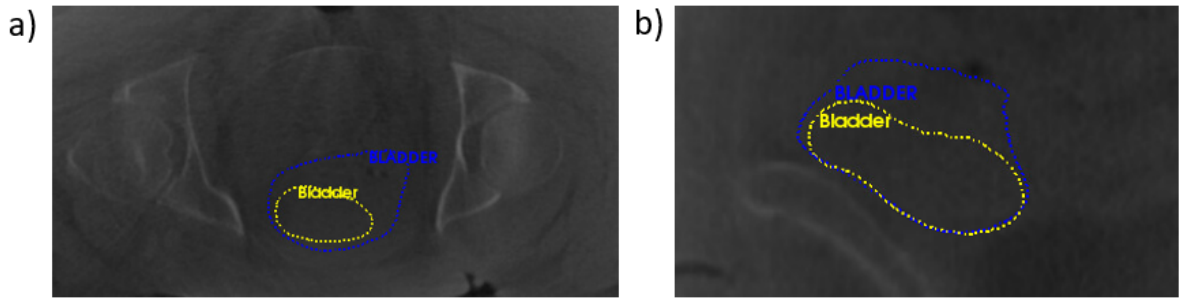


Figure 6: Bladder contours comparison from a patient in the dataset used for this project. The GS contour is shown in yellow, and the AG contour in blue. This AG contour volume was initially labeled with the score 2 in the scale from 1 to 5. (a) Axial view of the CBCT scan. (b) Sagittal view of the CBCT scan.

In Figure 7 two representative cases of a GS contour and an AG contour of the rectum can be observed. This AG contour was initially labeled with score 4, hence, classifying it as a good quality contour. The slices shown in this figure depict the maximal difference between the AG and GS contours found for this case in particular. This means that through all the slices except in the ones shown here, both contours were consistently keeping a good similarity. It is important to recall that the labeling was done considering overall delineation quality of the whole contour volume.

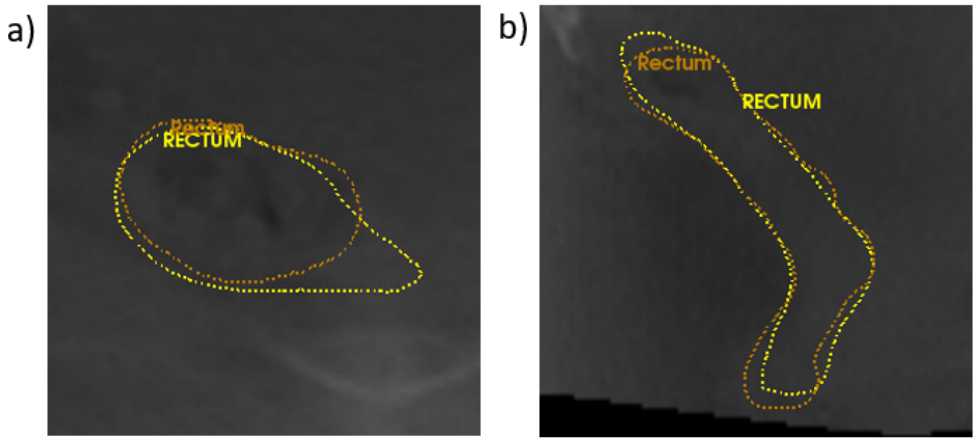


Figure 7: Rectum contours comparison from a patient in the dataset used for this project. GS contour is shown in brown, and the AG contour in yellow. This AG contour volume was initially labeled with the score 4 in the scale from 1 to 5. (a) Axial view of the CBCT scan. (b) Sagittal view of the CBCT scan.

3.2. Data preparation

For each contour volume, three subregions were defined: core region, inner shell and outer shell. The core region includes all the voxels inside the 3D contour itself, the inner shell is obtained from erosion of the core region, and the outer shell is obtained doing dilation of the original contour. Same amount of erosion and dilation was performed by defining a vector of the same size but used in opposite directions. The optimal size of the shells has not been studied, just an arbitrary shell size has been chosen according to the literature. These subregions were defined for taking into account not only the contour area itself, but also the surroundings, since a contour is perceived as an edge, which means that a noticeable change in pixel intensity has occurred. Therefore, the voxels surrounding the contour are also important to detect it. Masks were generated for each OAR (bladder and rectum) and for each of the previously described contour regions. A value of 0 is assigned to the pixels in the background, and a value equal to 1 is given to the pixels belonging to these structures.

To prepare the data for the feature extraction step described in the section 3.3, the mask volumes and their corresponding CBCT scans are needed in NIfTI format. *NiBabel* python package was used for this purpose.

3.3. Feature extraction and pre-processing of feature data

Medical images hide a lot of information that does not meet the eye and from which valuable clinical data can be obtained to improve patient care. Radiomics is the approach through which quantification of the phenotypic characteristics of the structures shown in these medical images is possible using advanced mathematical analysis [48]. It enhances the already existing medical data available for clinicians by creating imaging-based biomarkers which are non-invasive, and represent the biological properties of a structure in the image by extracting its radiomic features [49].

The contour QA methodology developed and implemented in this project is based on features obtained from the CBCT images. The features studied are not only the ones that can be perceived by the human eye, but also the ones that go beyond our perception. Studying also these features allows us to obtain very valuable information about the spatial arrangement of pixel intensities and their interrelationships, and how this is related to defining the structures shown in the CBCT scans. Multiple features can be extracted from medical images analysing different aspects of the structures involved. *PyRadiomics* is an open-source python package that was developed with the purpose of standardising radiomic analysis of medical images by providing a set of radiomic features as a benchmark for medical imaging analysis [49].

In this project, *PyRadiomics* has been implemented for feature extraction. First, we need to instantiate the extractor, which is the main building block for extracting the features. The extractor takes as input the CBCT images and their corresponding contour mask volumes, both in NIfTI format. To be

sure that the the correct pairs of contours and CBCT images were used, the NIfTI files were saved with filenames including the Patient ID, and the acquisition time in the case of the CBCT scans. For the case of the contour mask volumes, the OAR and the region of the contour mask (core, inner or outer shell) were also included in the filename.

The feature classes that were extracted from the given data include first-order statistics, shape-based, and texture-based features [49]:

- **First order statistics**: use commonly known metrics to give a description of the distribution of voxel intensities inside the image region defined by the given mask. These are generally histogram-based descriptors, hence the spatial relationships between voxels in the region of interest is lost. In this project, 18 features of this class have been extracted with *PyRadiomics* including metrics like mean, median, maximum and minimum intensity, entropy (describes how uniform or random is the distribution of the image intensities), kurtosis and skewness describing the histogram's flatness and asymmetry respectively [50].
- **Shape features**: include two- (2D) and three-dimensional (3D) descriptors for the physical appearance of the OARs, meaning size and shape. Some examples of these descriptors are the mesh volume, surface area, voxel volume or the surface to volume ratio among others. 14 features of this class were obtained.
- **Texture-based features** quantify the spatial relationships between voxel intensities. Two textural matrices were computed in this project from which multiple descriptors were obtained (see Figure 8) :

  Gray-level co-occurrence matrix (GLCM): evaluates spatial relationships of voxels by evaluating them in pair of values and assessing the probability of observing them at a given distance direction [51].

  Gray-level run-length matrix (GLRLM): for any voxel value, it counts the number of consecutive voxels aligned in a certain direction given that they have the same value [51].

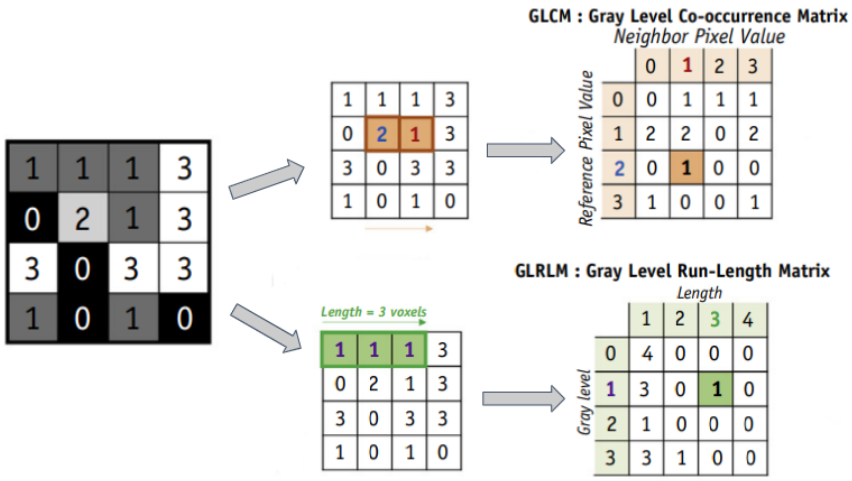


Figure 8: Graphic representation that shows how the GLCM and GLRLM are computed from the original gray-level image matrix of the ROI being evaluated [52]. In this particular example, the GLCM is looking into how many times the pair of gray-values 2 and 1 appear as neighbors in the original image. The GLRLM is showing the example of how many consecutive counts can be done of the gray-level 1 following the indicated direction.

Therefore, a total of 72 features (see Appendix A) were extracted from each contour region and saved in separate csv files with their corresponding identification filename including: patient ID, OAR, contour region, and acquisition date.

Then, Matlab was used for going through all these csv files, including bladder and rectum contours, and putting them all together in one single matrix containing the feature data of each contour mask volume per organ. Therefore, at the end of this step of the workflow, there were two matrices (saved

as csv files): one for the bladder and another one for the rectum, with 72 columns, one per feature. These matrices had as many rows as contours for each OAR present in the given dataset, multiplied by 3, due to the different subregions previously mentioned.

Finally, another script was done to add a last column to these matrices indicating the score label corresponding to each contour, as previously explained in sections 3.1.1 and 3.1.2. The feature data corresponding to the subregions of the contours were given the same label as the original contour from which they were defined. Therefore, the final feature data matrix for each OAR has 73 columns. The 73th column is the indicator of quality, hence it is the output that we will look for with the supervised ML methodologies introduced in section 3.4.

An additional preprocessing step required the relabeling of the contours. As previously mentioned, AG contours are labeled with a score from 1 to 5 depending on their quality. However, the clinical problem concerning our research question asks for only two classes of contours: accurate or inaccurate. In the clinical setup, medical physicists want to know if the AG contour is good enough or not for plan selection, not intermediate cases. Therefore, before starting the classification task, all the contours, AG and GS, were relabeled as follows: contours labeled with 1 or 2 were relabeled to 0, and contours labeled with a 4 or a 5 were assigned the label 1. AG contours initially labeled with a 3 were removed to keep only the extreme cases, simplifying the learning task for the classification algorithms avoiding ambiguity in the dataset. Label 3 represents the contours that are not good enough or not too bad, therefore, for initially training the algorithms, it is better to start with a binary classification problem and see how they behave with the given data.

3.4. Methodologies for automatic contour QA

Two different approaches were used for solving the binary classification problem that gives an answer to the research question. The fundamentals for both methodologies entail supervised ML, and these are the Random Forest classifier and Logistic Regression. Section 3.4.1 gives an explanation about the preliminary steps followed to prepare the data before using it as input of the RF. Then, the RF classifier is explained. In Section 3.4.2, the univariate and multivariate analysis performed using logistic regression are explained, showing how they are useful for giving an output for this binary classification problem.

Before implementing these methodologies, the original data for the bladder and the rectum were split in a training set and a test set. The ML model uses the training data to learn statistical relationships between the features and the output label. Once the model is trained, it is necessary to evaluate its performance on a different dataset (test set) and evaluate if, from the relationships learned from the training data, it is able to generalize to new unseen data. In this project, 80% of the data was used as the training set, and 20% as the test set. A stratified splitting of the data is performed to avoid class imbalance and ensure that the class labels are proportionally distributed between both datasets.

Python package *scikit-learn* has been used at different steps of the project to perform diverse tasks: preprocessing of the data (train/test split, standardization), dimensionality reduction, classification, and model selection and evaluation [53].

3.4.1. Random Forest Classifier

Random Forest (RF) is a supervised ML model that consists of the ensemble of multiple decision trees [54]. RF was chosen over decision tree because the latter is more prone to overfitting the training data and has higher variance. If the decision tree is not properly pruned, it could create more splits and grow more leaves meaning that it would adjust its predictions to every single input case (overfitting). This is why decision trees should not be very deep, which in the end limits their variance. Random Forests take the average of multiple non-correlated decision trees so that the variance is reduced, and the model is less prone to overfitting. This is achieved by randomly splitting the training set into subsets with different features obtaining predictions based on the sub-sample of features evaluated by each decision tree. Then, through majority voting, the class for each observation is decided.

The Random Forest classifier needs the data to be prepared in a certain way, for this reason, the preliminary steps are explained below. Moreover, RF has multiple parameters to which a default value is assigned unless they are specified. To ensure that the RF model used in this project is the most appropriate one for our dataset and classification problem, hyper-parameter tuning and optimization is performed. All this is explained in the next sections.

Preliminary steps

Each of the 72 features quantify different aspects of the images, hence each of them can be in a different range. Before using them as input of the ML model, it is necessary that the variance of the features are in the same range, centered around zero. The goal is to prevent the model from mistakenly learn that a feature is more important than another one just because its variance is orders of magnitude larger than the second feature's variance. Therefore, it is necessary to standardize the feature data achieving zero mean and unit variance. For this purpose, the *StandardScaler* function from *scikit-learn* was used before performing dimensionality reduction. The scaler was fit to the training data and then it was used to transform both, the training and the test set. Equation 9 shows that given the sample *x*, the formula for calculating its standard score is:

$$z = \frac{x - u}{s} \tag{9}$$

where *u* is the mean of the samples in the training set, and *s* is their standard deviation. By fitting the scaler to the training set, the mean and the standard deviation are computed and stored to be used for later scaling other datasets. Then, the training and the test data are transformed by performing centering and scaling independently on each feature.

After the standardization step, all the features are transformed to the same scale. However, some of these 72 features may be correlated and contain redundant information, hence, it is necessary to reduce the feature space and find the most representative variables. For this purpose, Principal Component Analysis (PCA) was performed. As previously explained in Chapter 2, PCA is a dimensionality reduction technique that is used to reduce the feature space by identifying the variables that are not very relevant while preserving as much information as possible. The output of this step are new variables that are obtained from linear combinations of the initial features. These combinations are called principal components, and they are uncorrelated and contain most of the information from the input features. They represent directions explaining a maximum amount of variance in the dataset, and the larger the variance, the larger the dispersion of the data points along the line, hence the more information this direction has. The first component has the largest variance, hence it captures the highest amount of information. The second principal component accounts for the next highest variance and is perpendicular (uncorrelated) to the first component, and so on [55]. The drawback of these principal components is that they don't have any real meaning because they are built from linear combinations, making their interpretability more difficult.

In this project, the parameters of the PCA were set up to find the principal components that explained 99% of the variance in the dataset.

Hyper-parameter optimization

After obtaining the principal components performing PCA in the training set, the data is ready to be used to train the RF model. Before this, the RF classifier has several parameters that should be tuned to find the most optimal model for our dataset and our classification problem. With this purpose, hyper-parameter tuning of the RF "settings" is performed using *GridSearchCV*, which is a tool from *scikit-learn* used for optimization of an estimator, in our case the RF classifier, and using a given set of parameters which are specified below:

- *n_estimators*: indicates the number of trees in the forest. The values given are [30, 50, 70, 100]. If it is not specified, the default value is 100, which for some datasets can be too many and lead to overfitting the training data. This is why it is important to tune this parameter and find the optimal value for our specific problem.
- *max_depth*: indicates the maximum depth of the tree. The default value is *None*, which means that the expansions of the tree continues until all leaves are pure or until they contain less than *min_samples_split* observations. A subset is pure when it contains samples from one single class. The values given to this parameter are [5, 8, 10, 15, 20].
- *min_samples_split*: this is the minimum number of samples required to split an internal node. If it is not specified, the default value is 2. The chosen values for finding the most optimal one are [2, 3, 5, 10].
- *min_samples_leaf*: this parameter is used to specify the minimum number of samples at a leaf node. For a better understanding of the terms used until now, the decision tree classification methodology works by splitting the data starting in the root node, then the internal node, and

finally the leaf nodes (the leaves represent the classes). The chosen values for *min_samples_leaf* in the project are [2, 3, 4, 5]. For reference, the default value is 1.

The values for these parameters were chosen in a wide and varied range that allows us to explore reasonable options without falling into overfitting or underfitting of the training data. Overfitting occurs when the model shows a good performance on the training data, but its generalization skills to other data are poor. Underfitting shows a poor performance on the training data and it also shows poor generalization to other data.

Moreover, to prevent these situations from occurring, *GridSearchCV* performs the search of the best model parameters doing a stratified k fold cross-validation (CV). This cross-validation is similar to a normal k fold CV in the sense that since it is stratified, the folds are created preserving a balanced distribution of samples for each class. In our case, 5 folds (k = 5) were used, which means that the dataset is split into 5 groups: 4 of them are used for training and the remaining is for testing. This changes until all the groups have been used as the test set.

Implementation of the final classification model

The RF model was built using the best parameters resulting from the hyper-parameter optimization. Then, by fitting the classifier model to the training data, the trained model was obtained, which was used on the test data to obtain predictions. Finally, the model performance on new unseen data was assessed using evaluation metrics as explained later in section 3.5.

3.4.2. Logistic Regression

Logistic regression (LR) comes from the statistics field and is considered a supervised classification algorithm [56]. It predicts a categorical output (y) given a predictor or input value (x). In the simplest case, the output variable only has two possible values, 0 or 1. Its name comes from the logistic (logit) function, which maps probabilities (values in the range [0,1]) to the full range of real numbers. The inverse of the logit function is known as the sigmoid function, and it is the base for understanding logistic regression [57].

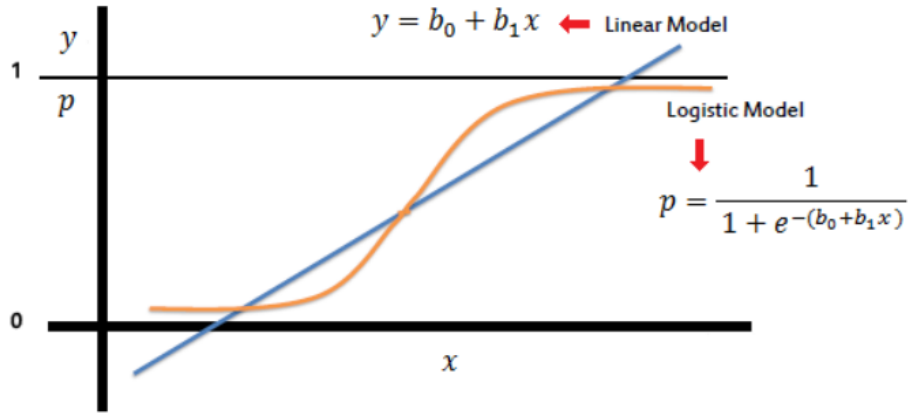


Figure 9: Comparison between linear regression and logistic regression models [58].

The sigmoid function is described as an S-shaped curve that can take any real number and map it into a value between 0 and 1, but never reaching the limits of the range (see Figure 9). The larger the value, the closer to 1 it will be mapped. To better understand its mathematical expression, first we have to look at the formula for a simple linear regression. The expression shown below corresponds to the case of univariate linear regression:

$$y = b_0 + b_1 * x \tag{10}$$

where b_0 is the intercept, and b_1 is the weight or coefficient that belongs to the relationship between the output (y) and the predictor or input values (x).

In the case of multivariate regression the formula would be:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n \tag{11}$$

Linear regression can't be used for the binary classification problem in this project, since the output that we are looking for is a categorical dependent variable (the class label: 0 or 1), and linear regression is usually used to predict the value of a continuous dependent variable (a numeric value). For that reason, logistic regression has been implemented. However, understanding the mathematical expression for linear regression is necessary to continue with the explanation of logistic regression. For simplicity, the remaining formulas will be built around the univariate linear regression case.

What logistic regression exactly does is to predict probabilities. It models the probability of the positive class (label 1) or endpoint. This means that its output is the predicted probability of a given predictor (x) of reaching the endpoint: P(y = 1|x) = p. Therefore, the probability of getting the negative class (0) is P(y = 0) = 1 - p. Recalling that the logistic regression formula is based on the logit function, we have the following expression:

$$logit = log\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right) \tag{12}$$

From this formula we can derive the expression for the odds ratio (OR), which is defined as the probability of an event happening divided by the probability of that event not happening. Another definition is that per unit increment of x, the increase of the chance of reaching the positive class (endpoint) is represented by the OR. The mathematical expression for OR is the ratio between brackets in Equation 12.

Now the mathematical expression for the previously mentioned sigmoid function can be obtained by setting Equation 10 equal to the logit function (Equation 12) resulting in the following expression:

$$P(y=1) = \frac{1}{1+e^{-(b_0+b_1*x)}} \tag{13}$$

where P(y = 1) is the probability of getting a predicted output value reaching class label 1. In the logistic regression model the constant (b_0) moves the curve left and right, while the slope (b_1) defines the steepness of the curve (see Figure 9).

Moreover, Equation 13 corresponds to the case of univariate logistic regression, but if we want to see how multivariate logistic regression formula looks like, we just need to plug in the corresponding extra variables:

$$P(y=1) = \frac{1}{1+e^{-(b_0+b_1*x_1+b_2*x_2+\ldots+b_n*x_n)}} \tag{14}$$

In this project, the methodology followed for implementation of the logistic regression model is based on the literature found [59, 60]. In these studies, first a univariate analysis of the data was done, and then the most predictive variables were obtained performing a multivariate logistic regression analysis with forward feature selection.

Univariate Feature Selection Analysis

Univariate feature selection considers each feature independently of each other and selects the ones that best represent the dataset by performing univariate statistical analysis. Before starting the analysis, the minimum and the maximum values were obtained for each feature in the training set of each OAR. Features with very extreme maximum and minimum values were rescaled to get the adjusted ORs, and they were different for the bladder and the rectum. The same features in the test set were rescaled accordingly.

After the pre-processing of the data, *Statsmodel* python library was used to fit a logistic regression model using maximum likelihood to the training data, with the 72 features as the independent variables, and the class label as the dependent variable. Then, using this model, statistical data was obtained. This data included adjusted Odds ratio (OR), 95% Confidence Interval (CI), and the p-value, which was obtained to evaluate the significance of the features. Moreover, to give some context to this statistical data, the median and the Interquartile Range (IQR) for each feature were obtained.

To start with the univariate feature selection process, the features with a p-value < 0.01 were considered significant and the rest were removed. However, for both OARs, 71 features were significant and only one had a p-value > 0.01, hence it was discarded. For both OARs, this feature was the *original_firstorder_Maximum*, which represents the maximum gray level intensity within the ROI.

Since most of the features were statistically significant, a correlation analysis was performed to identify the features giving repetitive information. For this purpose, the correlation matrix of all the features was obtained (see Appendix B). From here, the strongly correlated pairs of features were identified by setting the condition of having a Pearson's correlation coefficient > 0.8, following the methodology implemented in the study previously mentioned by Christianen et al. [59].

Moreover, the Area Under the Curve (AUC) was obtained for each of the features in the training set to have a quantitative representation of the performance of the logistic regression model considering each feature individually.

Using the list of strongly correlated feature pairs and the AUC scores, the least predictive features were identified and removed. This was done by identifying the feature in each pair with the lowest AUC and deleting it from the final set of features. The final outcome of this univariate feature selection step is a set of significant and non-correlated features based on the performance of a logistic regression model on the training set for each individual feature.

Multivariate Feature Selection Analysis

The reduced set of features obtained from the univariate analysis was used as input for multivariate analysis with forward feature selection (FFS). In forward selection, the variables are progressively added to the model, and in each step the selected feature is the one that increases the most the model's accuracy [61]. There are multiple methods for parameter selection, but FFS was chosen following the methodology implemented in the already mentioned studies used as reference [59, 60].

Final Logistic Regression model

Once the final set of features is selected after the multivariate analysis, the training and the test set are transformed by only keeping the columns corresponding to these variables. The final logistic regression model from *scikit-learn* is fit to the training data. The trained model is used to obtain predictions on the test set, which are used for evaluating the model performance, explained in the next section.

3.5. Evaluation of model performance

The resulting predictions obtained from the trained model on the test data were used for model performance evaluation, and for assessing the classification skills of the ML models. For classification problems, the evaluation metrics compare the expected class label from the test set to the predicted class label, or they interpret the predicted probabilities for the class labels for the classification problem in particular. The metrics used are explained below:

Accuracy

The accuracy is defined as the number of observations that were correctly predicted over the total number of observations. Its mathematical description is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

where TP indicates the true positives, which are the the good quality contours correctly classified as such; TN are the true negatives, indicating the bad quality contours correctly identified; FP are the false positives, which are the bad quality contours incorrectly classified as accurate (good quality) contours; lastly FN represents the false negatives, which are the good quality contours incorrectly labeled as inaccurate (bad quality) contours.

It is necessary to point out that the accuracy is affected by class imbalance in the dataset. In this project, the dataset for the bladder contained 42% of bad quality contours and 58% of good quality contours, while the rectum contained 37% and 63% respectively. Considering this situation, and specially for the rectum, the problem of class imbalance affects the dataset used in this project. Therefore, to ensure a good evaluation of the ML models' performance, more metrics should be considered.

Confusion Matrix
The confusion matrix (see Table 1) provides more insights into which classes are being predicted correctly, incorrectly, and the types of prediction errors that are happening. The default threshold probability set by *scikit-learn* to compute these evaluation metrics is 0.5, hence this is the cut-off probability value set for the binary classification results presented in Chapter 4.

Table 1: Schematic for the confusion matrix of a binary classification problem. The positive and negative classes are the true labels: positive class refers to the good quality contours (label 1); negative class refers to the bad quality contours (label 0)

	Positive Prediction	Negative Prediction
Positive Class	TP	FN
Negative Class	FP	TN

Sensitivity
It indicates the classifier's ability to identify positive labels. It refers to the true positive rate (TPR), and it is computed as shown below:

$$Sensitivity = \frac{TP}{TP + FN} \tag{16}$$

Specificity
It is the complement to sensitivity and describes the ability of the classifier of identifying negative samples. It is also called the true negative rate (TNR), and it is computed as follows:

$$Specificity = \frac{TN}{TN + FP} \tag{17}$$

The specificity is also defined as (1 - False Positive Rate (FPR)). The FPR is defined below.

Receiver Operating Characteristic (ROC) curve
The ROC curve evaluates how effective is a binary classifier at discriminating classes. It shows the behavior of a predictive model by plotting the true positive rate against the false positive rate (FPR) for a set of predictions. As previously mentioned, TPR is the same as sensitivity. The mathematical formula for FPR is as follow:

$$FPR = \frac{FP}{FP + TN} \tag{18}$$

The model gives a set of predictions under different probability thresholds. Each of these classification thresholds is a point on the plot and they are connected to form the ROC curve. Figure 10 shows classifiers with different skills represented by their corresponding ROC curve. A model with no classification skills (e.g. predictions for the positive class (label 1) are under all the thresholds) is represented by the diagonal line. A perfect classifier would be at the top left of the plot, while a classifier that is worse than no skill would be below the diagonal line.

Area under the curve (AUC)
This metric is complementary to the ROC curve and as its name indicates, it summarizes in one single number the performance of the classifier being evaluated by quantifying the area under the ROC curve. The AUC metric only cares about the ability of the classifier to separate the two classes in the dataset, and its value ranges from 0 to 1. Following the example given with the ROC curve, a no skill classifier would have an AUC score of 0.5, because at least half of the predictions will be correct, and all the other half will not be since all predictions for the positive class will be under all the thresholds as previously mentioned. A perfect classifier will have an AUC score of 1, because all predictions match the true label, meaning that all observations are correctly classified.

Figure 11 shows an example of the visual representation of the predictions obtained from a logistic regression model. AUC can be interpreted as the probability that the model ranks a positive sample higher than a negative one. Looking at Figure 11, AUC indicates the probability that a random positive observation (green) is positioned to the right of a random negative one (red) [63].

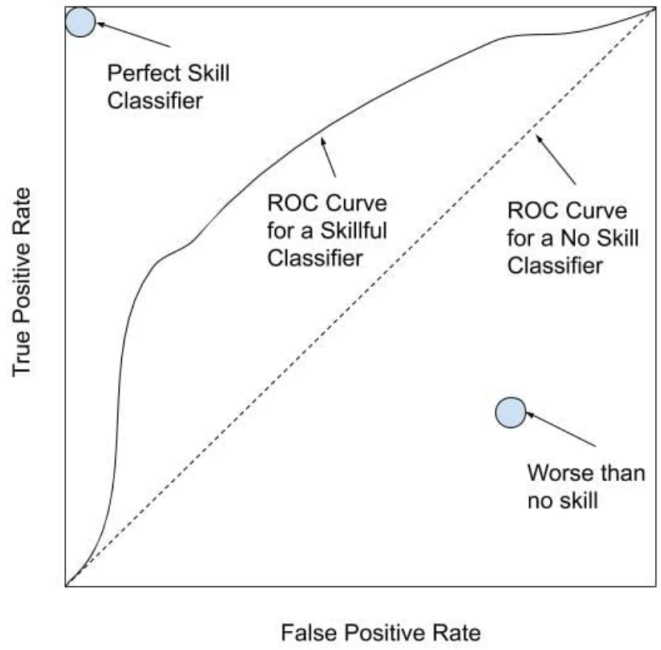


Figure 10: ROC curve plot showing the meaning of different curves related to the skills of the corresponding classification model [62].

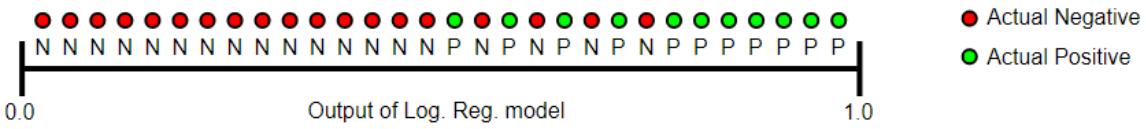


Figure 11: Predictions from a logistic regression model ranked in ascending order of prediction probabilities [63].

It is necessary to take into account that AUC is not affected by the classification threshold. Its value measures the quality of the model's predictions regardless of the chosen threshold. Moreover it is scale-invariant, meaning that it doesn't look at the absolute values of the predictions, it is more a measure of how well the predictions are ranked. If the predictions are transformed but the relative ranking of predictions is preserved, there is no impact on the AUC value.

4

Results

The results obtained for the bladder and the rectum with the two methodologies explained are shown separately in the next sections. As explained in Chapter 3, this project faces a binary classification problem and the results obtained for each OAR with the random forest classifier and the logistic regression model show each model's performance according to the two classes considered: accurate (label 1) or inaccurate (label 0) contours.

It is important to take into account that scikit-learn obtains the predictions for the binary classification by setting a default threshold of 0.5, meaning that if a predicted probability is in the range [0, 0.49] it will be classified as a bad quality contour (label 0). Therefore, if the predicted probability is in the range [0.5, 1], the assigned class will be label 1, i.e. good quality contour. Table 4 shows a summary of all the results for the bladder and the rectum.

4.1. Bladder results

4.1.1. Random Forest Classifier

Dimensionality reduction with PCA on the bladder training data resulted in 21 principal components that explain 99% of the the variance. The best Random Forest classifier for the training feature data of the bladder after dimensionality reduction was built based on the results obtained from the hyperparameter optimization. The optimized value of these parameters for the given training data were:

- max_depth = 15
- min_samples_leaf = 2
- min_samples_split = 3
- n_estimators = 100

After training the model with the mentioned parameters, predictions were obtained on the test data with an accuracy of 0.807. The performance of the model is shown in Figure 12, where it can be observed that it performed well on the test data as it is indicated by the high AUC value.

When obtaining the confusion matrix on the test data (399 observations in total) the following results were obtained:

- TP = 63% (251/399)
- TN = 18% (71/399)
- FP = 15% (61/399)
- FN = 4% (16/399)

These results gave a sensitivity of 94% and a specificity of 54%.

4.1.2. Logistic Regression

Univariate Analysis

After performing univariate feature selection together with the correlation analysis, the selected features are shown in Table 2.

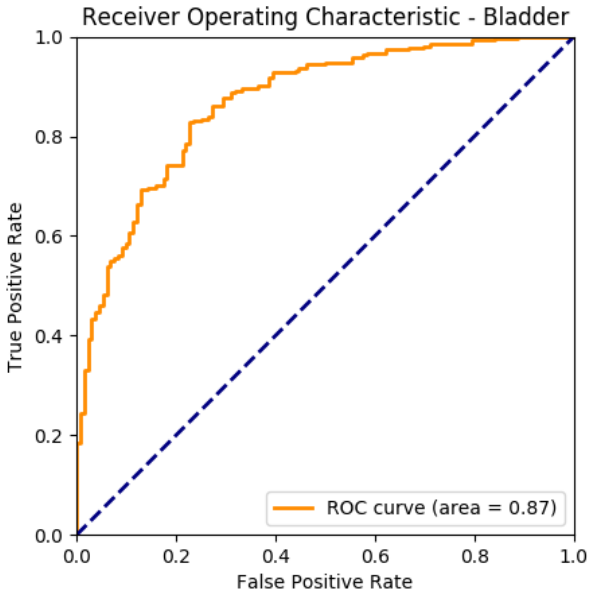


Figure 12: ROC Curve and AUC showing the random forest classifier performance for the bladder dataset.

Table 2: Selected features for the bladder resulting from the univariate analysis. The features in bold type are the final 13 features selected after multivariate analysis. *Idmn = Inverse difference moment normalized. Imc2 = informational measure of correlation 2.*

	Median	(IQR)	OR	95% CI	p-value
original_shape_Elongation	0.833	(0.764 - 0.887)	2.364	(2.082 - 2.683)	<0.01
original_shape_Maximum2DDiameterSlice	102.400	(91.831 - 112.700)	1.007	(1.006 - 1.008)	<0.01
original_shape_Sphericity	0.262	(0.225 - 0.784)	3.454	(2.786 - 4.283)	<0.01
original_firstorder_90Percentile	-312.000	(-382.500 - -246.000)	0.999	(0.998 - 0.999)	<0.01
original_firstorder_InterquartileRange	54.000	(38.000 - 78.000)	1.012	(1.010 - 1.014)	<0.01
original_firstorder_Kurtosis	3.272	(2.664 - 5.215)	1.096	(1.076 - 1.117)	<0.01
original_firstorder_Minimum	-525.000	(-604.000 - -484.000)	0.999	(0.999 - 0.999)	<0.01
original_firstorder_Range	339.000	(254.000 - 509.000)	1.002	(1.001 - 1.002)	<0.01
original_firstorder_Skewness	0.322	(-0.018 - 0.794)	1.662	(1.473 - 1.876)	<0.01
original_firstorder_TotalEnergy	1.601E+10	(9.752E+09 - 2.744E+10)	1.001	(1.001 - 1.001)	<0.01
original_glcm_ClusterShade	9.570	(-0.182 - 50.294)	1.004	(1.003 - 1.005)	<0.01
original_glcm_DifferenceAverage	0.522	(0.4269 - 0.660)	3.467	(2.893 - 4.153)	<0.01
original_glcm_DifferenceVariance	0.365	(0.282 - 0.633)	1.841	(1.596 - 2.123)	<0.01
original_glcm_Idmn	0.997	(0.996 - 0.998)	2.023	(1.822 - 2.246)	<0.01
original_glcm_Imc2	0.936	(0.883 - 0.964)	2.191	(1.953 - 2.457)	<0.01
original_glcm_InverseVariance	0.421	(0.388 - 0.459)	5.708	(4.451 - 7.318)	<0.01
original_glrlm_GrayLevelNonUniformity	10070.000	(6514.600 - 15636.000)	1.000	(1.000 - 1.000)	<0.01
original_glrlm_LongRunHighGrayLevelEmphasis	319.800	(225.740 - 492.490)	1.001	(1.001 - 1.001)	<0.01
original_glrlm_LongRunLowGrayLevelEmphasis	0.160	(0.073 - 0.296)	3.077	(2.173 - 4.358)	<0.01
original_glrlm_RunLengthNonUniformity	24508.000	(16098.000 - 36130.000)	1.000	(1.000 - 1.000)	<0.01
original_glrlm_ShortRunHighGrayLevelEmphasis	33.413	(20.667 - 59.307)	1.009	(1.007 - 1.011)	<0.01
original_glrlm_ShortRunLowGrayLevelEmphasis	0.017	(0.010 - 0.024)	1.225	(1.168 - 1.285)	<0.01

Multivariate Analysis

In Figure 13 it can be observed that for the bladder feature data, the highest accuracy value corresponds to 13 features out of 22 features resulting from the previous univariate analysis step. These 13 features are 4 histogram-based, 6 texture-based features related to the GLCM, and another 3 texture-based features related to GLRLM. The name of the specific selected features are highlighted in Table 2.

Finally, these 13 features are used for training the final logistic regression model. An accuracy of 0.747 was obtained when evaluating the predictions of the model. Figure 14 shows the performance of the model on the test data. The AUC value of 0.77 shows that the performance of the logistic regression classifier for the bladder data is worse compared to the random forest model.

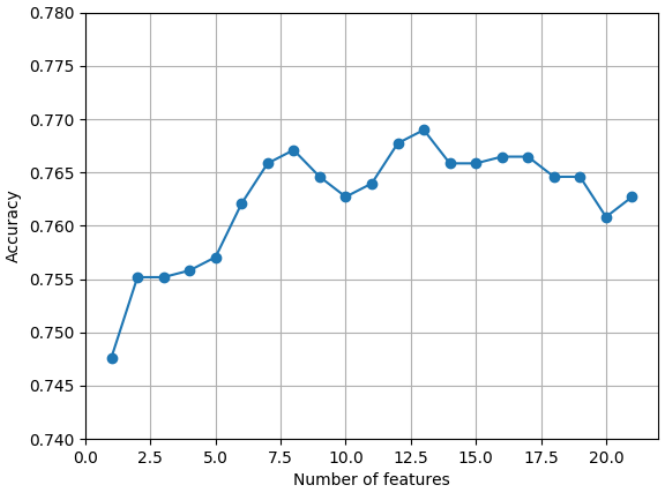


Figure 13: Plot showing how the accuracy varies during the Forward Feature Selection process for the bladder as more features are added one by one to the model.

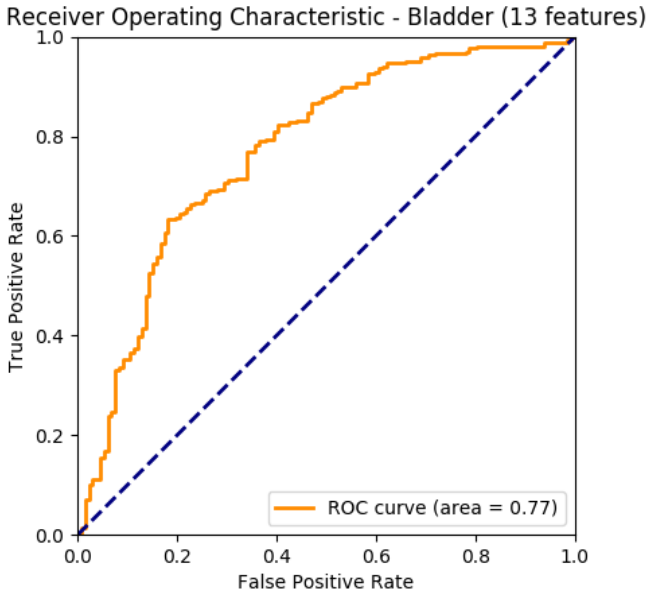


Figure 14: ROC curve and AUC showing the performance of the logistic regression model trained on the bladder dataset reduced to the selected 13 features resulting from the multivariate analysis.

The results obtained from the confusion matrix on the test data (399 observations in total) are shown below, resulting in a sensitivity of 91% and a specificity of 42%:

- TP = 61% (243/399)
- TN = 14% (55/399)
- FP = 19% (77/399)
- FN = 6% (24/399)

4.2. Rectum results

4.2.1. Random Forest Classifier

Dimensionality reduction with PCA on the rectum training data resulted in 23 principal components that explained 99% of the variance. The best RF model for the rectum data after PCA was built based on the best parameters obtained from hyper-parameter optimization. The values for each of these parameters are:

- max_depth = 20
- min_samples_leaf = 4
- min_samples_split = 3
- n_estimators = 70

After training the model with these parameters and obtaining predictions on the test data, an accuracy of 0.802 was obtained. The performance of the model is shown with the ROC curve in Figure 15, and as it happened in the case of the bladder, the high AUC value indicates that the model performed well on the new unseen data.

The results obtained from the confusion matrix on the test data (353 observations in total) showing the quality of the output of the classifier are shown below:

- TP = 70% (246/353)
- TN = 10% (37/353)
- FP = 17% (59/353)
- FN = 3% (11/353)

These results gave a sensitivity of 96% and a specificity of 38%.

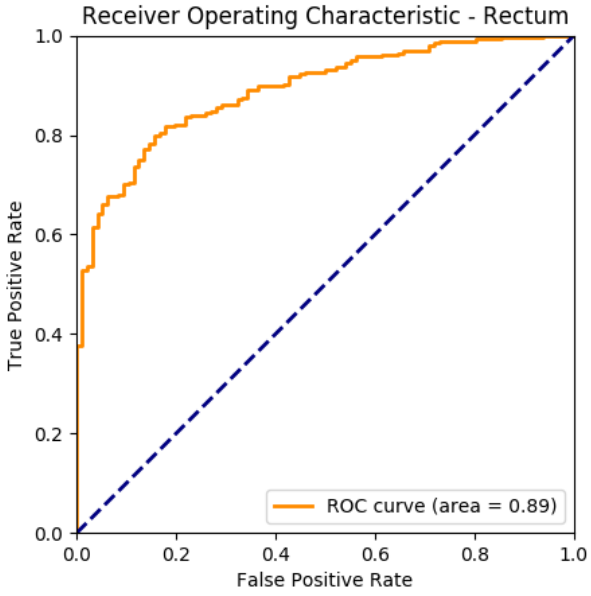


Figure 15: ROC Curve and AUC showing the random forest classifier performance for the rectum dataset.

4.2.2. Logistic Regression

Univariate Analysis

A statistical analysis of the feature data was performed through a correlation analysis and univariate feature selection. The results for this are shown in Table 3.

Table 3: Selected features for the rectum resulting from the univariate analysis. The features in bold type are the final 19 features selected after multivariate analysis. *Idmn = Inverse difference moment normalized.*

	Median	IQR	OR	95% CI	p-value
original_shape_Elongation	0.461	(0.391 - 0.553)	6.153	(4.851 - 7.804)	<0.01
original_shape_Flatness	0.357	(0.303 - 0.412)	10.692	(7.842 - 14.577)	<0.01
original_shape_MajorAxisLength	116.220	(106.600 - 125.230)	1.009	(1.008 - 1.010)	<0.01
original_shape_Maximum2DDiameterColumn	84.172	(70.364 - 97.864)	1.013	(1.011 - 1.014)	<0.01
original_shape_Maximum2DDiameterSlice	60.638	(50.220 - 71.642)	1.016	(1.014 - 1.018)	<0.01
original_shape_Sphericity	0.298	(0.265 - 0.640)	7.268	(5.534 - 9.544)	<0.01
original_shape_SurfaceArea	21840.000	(15059.000 - 30802.000)	1.000	(1.000 - 1.000)	<0.01
original_firstorder_10Percentile	-452.000	(-528.000 - -401.000)	0.998	(0.998 - 0.999)	<0.01
original_firstorder_90Percentile	-352.500	(-391.000 - -230.000)	0.998	(0.997 - 0.998)	<0.01
original_firstorder_InterquartileRange	52.000	(39.000 - 82.000)	1.011	(1.010 - 1.013)	<0.01
original_firstorder_Kurtosis	6.624	(3.929 - 11.550)	1.075	(1.062 - 1.088)	<0.01
original_firstorder_Range	589.500	(441.000 - 841.000)	1.001	(1.001 - 1.002)	<0.01
original_firstorder_Skewness	-1.187	(-2.002 - -0.180)	0.722	(0.677 - 0.770)	<0.01
original_firstorder_TotalEnergy	1.043E+10	(6.542E+09 - 1.544E+10)	1.005	(1.004 - 1.006)	<0.01
original_glcm_ClusterProminence	3380.300	(434.680 - 14902.000)	1.000	(1.000 - 1.000)	<0.01
original_glcm_DifferenceAverage	0.731	(0.559 - 0.977)	3.313	(2.862 - 3.836)	<0.01
original_glcm_DifferenceVariance	0.776	(0.471 - 1.386)	1.727	(1.578 - 1.891)	<0.01
original_glcm_Idmn	0.997	(0.997 - 0.998)	2.683	(2.386 - 3.018)	<0.01
original_glcm_InverseVariance	0.455	(0.421 - 0.479)	8.798	(6.774 - 11.429)	<0.01
original_glcm_SumAverage	28.515	(21.715 - 42.155)	1.031	(1.028 - 1.035)	<0.01
original_glrlm_GrayLevelNonUniformity	6064.900	(4604.100 - 7798.700)	1.000	(1.000 - 1.000)	<0.01
original_glrlm_LongRunHighGrayLevelEmphasis	933.480	(544.780 - 1599.700)	1.001	(1.001 - 1.001)	<0.01
original_glrlm_LongRunLowGrayLevelEmphasis	0.033	(0.013 - 0.062)	1.046	(1.034 - 1.059)	<0.01
original_glrlm_RunEntropy	4.616	(4.328 - 4.925)	1.234	(1.204 - 1.266)	<0.01
original_glrlm_RunLengthNonUniformity	18588.000	(12154.000 - 27458.000)	1.000	(1.000 - 1.000)	<0.01
original_glrlm_ShortRunLowGrayLevelEmphasis	0.005	(0.003 - 0.008)	1.838	(1.611 - 2.097)	<0.01

Multivariate Analysis

Figure 16 shows that for the rectum feature data, the highest accuracy value obtained with FFS corresponds to when the logistic regression model is fit to 19 features out of the 26 resulting from the univariate analysis. These features correspond to 6 shape class features, 4 histogram-based, 4 texture-based features corresponding to GLCM, and 5 texture-based features related to GLRLM. The name of these 19 selected features are shown in bold type letters in Table 3.

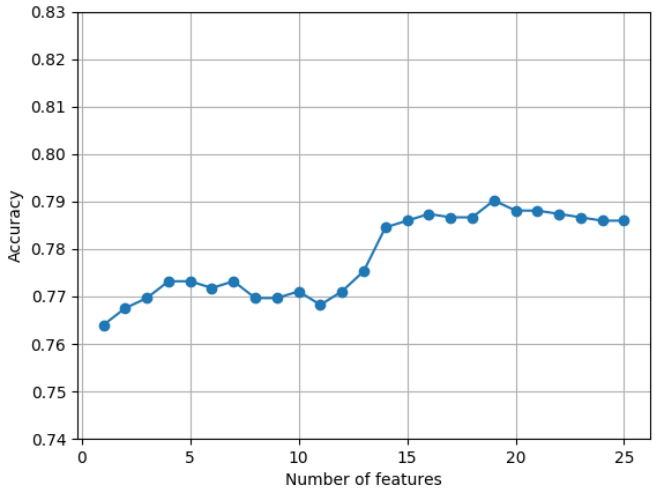


Figure 16: Plot showing how the accuracy varies during the Forward Feature Selection process for the rectum as more features are added one by one to the model.

The rectum dataset was reduced to 19 features and was used for training the final logistic regression model. An accuracy of 0.796 was obtained when testing the trained model on new unseen data. Figure 17 shows the performance of the model on the test data. The AUC value of 0.84 indicates that the model

performed well on the test data with only 19 features from the original 72.

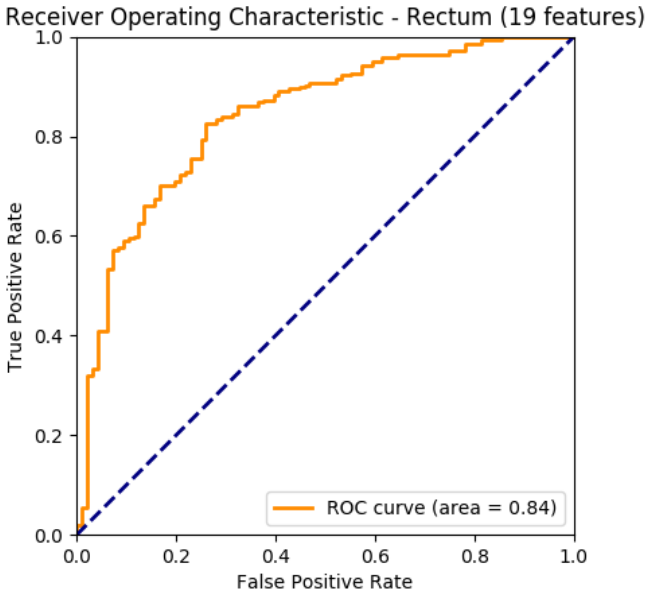


Figure 17: ROC curve and AUC showing the performance of the logistic regression model trained on the rectum dataset reduced to the selected 19 features resulting from the multivariate analysis.

The results obtained from the confusion matrix on the test data (353 observations in total) are shown below, resulting in a sensitivity of 95% and a specificity of 38%:

- TP = 69% (244/353)
- TN = 10% (37/353)
- FP = 17% (59/353)
- FN = 4% (13/353)

A summary of all the results described until now for the bladder and the rectum is shown in Table 4.

Table 4: Summary of the results.

	Bladder		**Rectum**	
	Random Forest	Logistic Regression	Random Forest	Logistic Regression
AUC	0.87	0.77	0.89	0.84
Sensitivity (%)	94	91	96	95
Specificity (%)	54	42	38	38

4.3. Blind Test

To test both methodologies implemented, and do a case-specific evaluation, a blind test of the random forest classifier and the multivariate logistic regression model was performed. A bladder and a rectum contour have been selected and tested in both algorithms, previously making sure that these contours were not present in the training set used to train the parameters of these classifiers. Three different cases have been evaluated per OAR: an AG contour initially classified as a bad quality contour (labels 1 or 2 in the initial dataset), an AG contour labeled as an accurate delineation (initially labeled with 4 or 5), and an AG contour classified with label 3. It's important to remember that all the contours belonging to this class were deleted from the dataset to avoid ambiguity and only keep extreme cases, reducing the possibility of creating confusion while training the classification algorithms.

4.3.1. Bladder: bad quality automatically-generated contour

Figure 18 shows the AG bladder contour of the patient randomly selected to perform this blind test. This AG contour was initially labeled with a 1 hence, it was relabeled to 0 for the binary classification problem. The fact that its initial label was 1 shows that this bladder contour has a very bad quality to be considered for plan selection assessment.

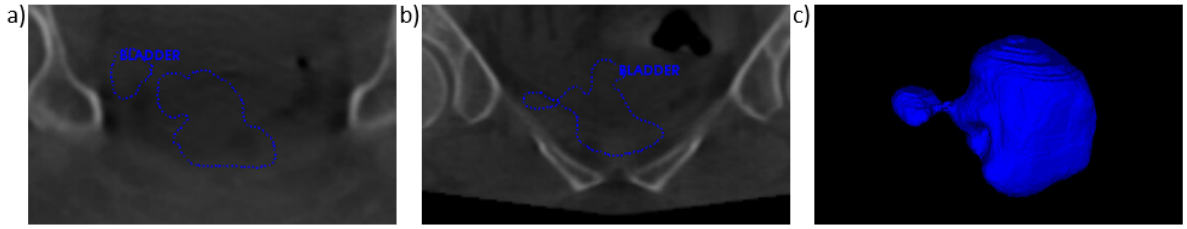


Figure 18: Automatically-generated bladder contour volume classified as with bad quality (label 0). (a) Axial view. (b) Coronal view. (c) Bladder contour volume rendering showing how the delineated bladder has a protuberance that does not belong to a normal bladder volume.

Random Forest classifier: using the previously trained model for the bladder data, the prediction for the class label obtained for this sample contour correctly classified it with label 0. Recall that when directly predicting the class label instead of the probabilities, *scikit-learn* will internally assign class labels using 0.5 as the default probability threshold. When looking into the predicted probabilities for each class, label 0 was predicted with a probability of 0.616 and label 1 with a probability of 0.384.

Logistic Regression model: for the same contour as the one evaluated with the RF classifier, using the Logistic Regression model obtained after univariate and multivariate feature selection the predicted class label was also 0. When looking into the probabilities, LR classifier was a bit more sure about the bad quality of the contour, compared to the RF model. The predicted probabilities were 0.673 for the class label 0, and the class label 1 was predicted with a probability of 0.327.

4.3.2. Bladder: good quality automatically-generated contour

Figure 19 shows the selected good quality AG contour for the bladder, initially labeled with a 4, and later relabeled to 1 for the binary classification problem. Therefore, this contour volume was considered as good enough for plan selection assessment.

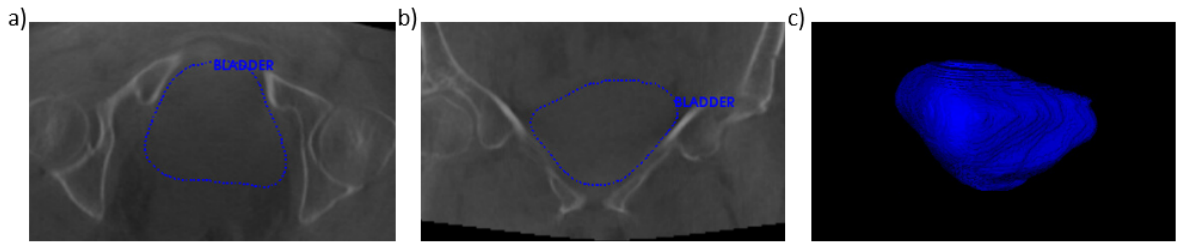


Figure 19: Automatically-generated bladder contour volume classified as accurate (label 1). (a) Axial view. (b) Coronal view. (c) Bladder contour volume rendering showing a shape closer to what is expected from a normal bladder volume.

Random Forest classifier: the model correctly classified the contour volume assigning the label 1. This class was predicted with a probability of 0.906, while the class label 0 got a predicted probability of 0.094.

Logistic Regression model: label 1 was correctly predicted for the contour with the default probability threshold of 0.5. When looking at the predicted probabilities, label 0 was predicted with a probability of 0.329, while a probability of 0.671 was assigned to label 1. The logistic regression model was less confident than the random forest classifier when assigning the class label 1 to the contour.

4.3.3. Bladder: label 3 automatically-generated contour

A contour labeled with score 3 in the initial dataset (see Figure 20), means that it is not good enough to be considered of good quality, but is not too bad either. When trying to assign a class label using models that were trained for a binary classification problem (characterized by extreme cases, i.e. negative/positive outcome) a noticeable difference has been observed in their predictions as explained below.

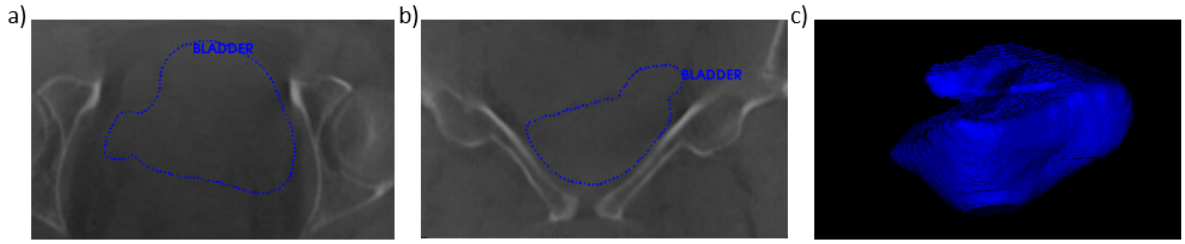


Figure 20: Automatically-generated bladder contour volume labelled with score 3. (a) Axial view. (b) Coronal view. (c) Bladder contour volume rendering.

Random Forest classifier: this model assigned the label 1 to the contour. When looking at the predicted probabilities, the RF model was more sure about the class label 1 than what it would be expected from a label 3 contour. The predicted probability for label 0 was 0.134, while label 1 was predicted with a probability of 0.866.

Logistic Regression model: class label 1 was also assigned to the contour. However, when looking at the predicted probabilities, the logistic regression model was less sure about assigning this class label. Label 0 was predicted with a probability of 0.357, and a predicted probability of 0.643 was obtained for label 1.

4.3.4. Rectum: bad quality automatically-generated rectum contours

In Figure 21 a rectum contour volume initially labeled with 1 and then relabeled to 0 can be observed. When using the features of this contour as input for getting predictions from the classification models, the results obtained were as follows:

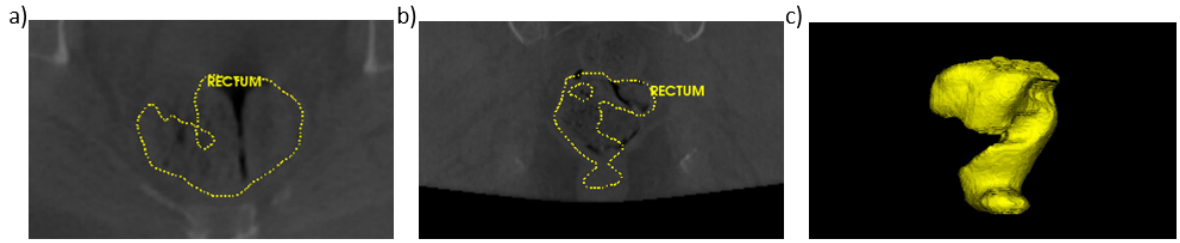


Figure 21: Automatically-generated rectum contour volume classified as inaccurate (label 0 in the binary classification). (a) Axial view. (b) Coronal view. (c) Rectum contour volume rendering showing an abnormal shape due to errors in the segmentation algorithm.

Random Forest classifier: in this case, the assigned class was incorrect, choosing the label 1 instead of 0. This is because the predicted probability for the good quality class was 0.680, which is bigger than the cutoff probability of 0.5. Label 0 was predicted with a probability of 0.320. Therefore, the contour is classified as accurate when in reality has very bad quality. In this example, it is noticeable the importance of changing the threshold to our interest for reducing the false positive cases. For the clinical problem concerning this project, is worse having false positives than false negatives (classifying an accurate contour as inaccurate). In clinical practice, is more risky for the patient to have false positives, because a bad quality contour would be classified as accurate which may end up influencing in a negative way the decision making process of the RTTs for plan selection.

Logistic Regression model: the contour was miss-classified as well by predicting the class label 1 instead of 0. The label 1 was predicted with a probability of 0.622, while the prediction probability of label 0 was 0.378.

4.3.5. Rectum: good quality automatically-generated rectum contours

The rectum contour selected is shown in Figure 22 and was initially labeled with a 4. The results obtained regarding predictions of the class are explained further below.

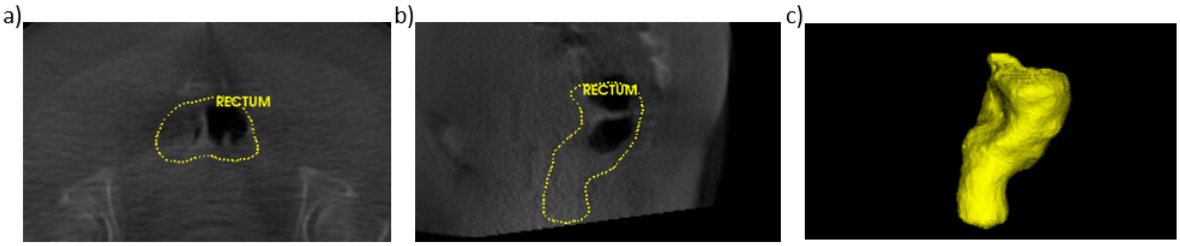


Figure 22: Automatically-generated rectum contour volume classified as accurate (label 1 in the binary classification). (a) Axial view. (b) Sagittal view. (c) Rectum contour volume rendering.

Random Forest Classifier: the contour was correctly labeled as a good quality contour, predicting the label 1. Regarding the probabilities of the predicted output, label 1 was predicted with a probability of 0.856.

Logistic Regression model: similar results as with the RF model were obtained. The rectum contour volume was classified as well as a good quality contour. The probability of being from the class with label 1 was 0.842, while the remaining 0.158 was the predicted probability for label 0.

4.3.6. Rectum: label 3 automatically-generated contour

Figure 23 shows an example of a rectum contour volume removed from the final dataset used in the project, since it was classified with label 3. The results obtained from testing both classifiers on this contour volume are described below.

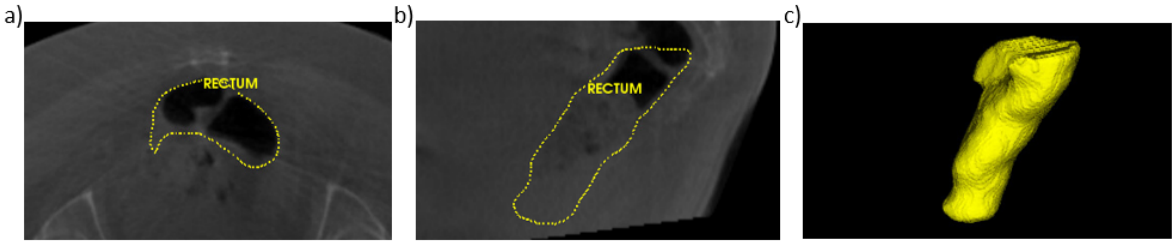


Figure 23: Automatically-generated rectum contour volume of class label 3. (a) Axial view. (b) Sagittal view. (c) Rectum contour volume rendering.

Random Forest classifier: the class label assigned to the rectum contour volume was label 1. The predicted probability for this class was 0.800, while the label 0 was predicted with a probability of 0.200, clearly indicating that for the RF classifier, this label 3 contour is a good quality delineation.

Logistic Regression model: the predicted class label was 1 as well. However, the predicted probabilities were more balanced, as it would be expected for a label 3 contour volume: the probability of getting label 0 as output was 0.361, while a probability of 0.639 was obtained for label 1.

5

Discussion

In this project, a comparison study of two supervised ML strategies (Random forest and Logistic regression) was performed by doing feature extraction and selection of quantitative image descriptors, including shape variables, histogram-based and texture-based features. This was done with the purpose of obtaining quantifiable characteristics of the contours and getting the statistical relationship of this data to their quality labels. These statistical relationships allow us to predict the class label of new contours.The ML models were evaluated to make sure that they learnt adequately from the training data, keeping their ability to generalize to new unseen data. Their performance was measured with the AUC value, and the quality of the predictions was evaluated with the sensitivity and specificity. The overall results indicate that the Random forest classifier has a better performance for both OARs evaluated. However, the results also show that the threshold probability for separating the contours in the two classes should be modified for achieving better specificity values.

5.1. Creation of subregions

Contour volumes were divided into core region, inner and outer shells taking as reference the study published by Zhang et al. [45]. They studied quantitative image features in the pancreas head and duodenum contours to classify them as either accurate or inaccurate delineations. Their final goal was to use the designed contour QA strategy as part of an auto-segmentation tool for online adaptive replanning. The pancreas head and duodenum are organs subjected to significant inter-fraction deformations, which is also a problem faced during LACC treatment but with the bladder and the rectum. The rationale behind defining these subregions is that only evaluating the contours themselves (core region) does not provide enough and relevant data about the contour characteristics to create a quality prediction model with them that is robust and consistent enough. Including the core region in the dataset allows us to check the overall position of the contour. Its feature values are expected to be more homogeneous, since it is the inside of the organ, and it is less affected by changes in textures at the surface of the organ as well as partial volume effects, which occur when a voxel contains information about more than one tissue type [64]. The inner and the outer shells provide information about the areas surrounding the contour. A contour is a boundary between the inside of the organ and the surrounding structures. Therefore, including these two regions for each contour in the dataset, adds the information of having more constant feature values within the OAR that start changing in the outer shell (beyond the contour volume) as more structures surrounding the OAR are included.

5.2. Selected features

5.2.1. Bladder

In the case of the bladder, after the first statistical analysis of the data with the univariate analysis, 3 shape features were selected (Table 2):

- *Elongation*: describes the relationship between the two largest principal components in the given region.

- *Maximum2DDiameterSlice*: indicates the largest pairwise Euclidean distance in the axial plane measured between the vertices of the surface mesh of the OAR volume.
- *Sphericity*: quantifies the roundness of the OAR with respect to a sphere.

Theoretically, the shape of the bladder can be assumed to be similar to an ellipsoid volume as shown in a previous study [65] hence, these shape features would provide valuable information. However, in the situation in which the bladder is considered in this project, it is subjected to constant deformations. Therefore, these shape descriptors are not consistent or robust enough to infer statistical relationships between their values and the corresponding class labels. In fact, this is confirmed by the selected features after multivariate logistic regression analysis, where all the shape features were discarded and only intensity-based features were kept i.e., histogram-based and texture-based.

5.2.2. Rectum

The univariate analysis of the rectum training data gave as a result 7 shape features as shown in Table 3. After FFS, 6 of these shape features were kept for the final implementation of the multivariate logistic regression model. This denotes the importance of shape descriptors in the case of the rectum, contrary to the bladder. Besides the elongation and maximum 2D diameter slice, described above, the other 4 selected features are indicated below:

- *Flatness*: describes the relationship between the smallest and the largest principal components in the given region.
- *MajorAxisLength*: considering that there is an ellipsoid encapsulating the ROI, this feature gives the largest axis length of this ellipsoid calculated using the major principal component.
- *Maximum2DDiameterColumn*: indicates the largest pairwise Euclidean distance in the coronal plane measured between the vertices of the surface mesh of the OAR volume.
- *SurfaceArea* of the total triangle mesh defining the volume of the OAR.

Sphericity was also selected after univariate analysis, however it was not relevant enough as a descriptor of the rectum' shape to be selected in the multivariate analysis, because as it can be expected, the shape of the rectum is not comparable to a sphere.

These results show the importance of the feature selection step, evaluating each organ independently and not assuming a set of features that work in every case. Previous literature show the common use of geometric and location-based features for assessing contour quality [40, 41]. However, the organs and tumors evaluated are usually in static regions, like head and neck. These descriptors worked well in this area but they are not robust enough to be used as a standard methodology for contour QA. Specially they are not consistent enough to be used as single descriptors for organs with highly variable position and shape. In fact, Zhang et al. [45] directly discarded geometric and shape features. They presented a methodology for automatic contour QA of the pancreas head and duodenum (subjected to large inter-fraction deformations) based on the feature extraction of only histogram-based, gray-level co-occurrence (GLC)-based and gray-level run-length (GLR)-based descriptors. They still achieved a great performance of their decision tree model with an average sensitivity and specificity of 85% and 91% respectively for the pancreas head contours, and 92%/92% respectively for the duodenum contours.

This shows that to go in the correct direction for achieving a standardized methodology for feature extraction and selection for automatic contour QA, texture-based features are non-negotiable. Shape features can be useful for some organs and not for others, as shown with the bladder and the rectum results. Therefore, they should also be included in the feature dataset to give the ML algorithm the chance to choose whether these features are relevant or not.

5.3. Model performance and classification skills

As shown in Table 4, the RF and LR models do not have major problems identifying the positive labels i.e., good quality contours. This is indicated by the high sensitivity values for all of them, above 90%. However, when looking at the ability of the classifiers to identify the bad quality contours, the specificity values are lower than desired. This is especially true for the LR model for the bladder data, and both classifiers for the rectum data, since their ability to detect bad quality contours is lower than 50%.

The research question presented in this project targets the clinical problem of plan selection. The AG contours obtained for supporting the decision-making process of choosing a treatment plan from the plan library, need to be checked to ensure their good quality. However, they do not need to be perfectly delineated contours; they only need to be good enough for aiding plan selection.

The AUC values in Table 4 indicate that the performance of the models is good since overall, they are around 80% (or higher) for both, the bladder and the rectum. Therefore, they have the potential to correctly classify the contours. However, more appropriate classification results for our goal would be achieved if the number of bad quality contours labeled as accurate (false positives) is decreased.

Figure 24 shows two different cases of cutoff points, A and B. It will be used as an illustrative example to better correlate the explanation given to its visual representation in a ROC curve. In this figure, cutoff point A has a bigger sensitivity value and lower specificity, which translates into more false positives. This means that with cutoff A the good quality contours can be accurately detected. However, the classifier struggles more to detect the bad quality contours, which leads to a higher misclassification of these contours. With cutoff B it is less likely that the classifier correctly detects the true positives (lower sensitivity). However, it is more probable that due to the higher specificity, the true negatives (inaccurate contours) are better identified.

Identifying the current situation of the classifiers in Table 4 (low specificities) with cutoff A, it is expected that they would show better performance with cutoff B (higher specificity), according to what is expected from the classifiers in this project. This means that for the clinical problem that the research question tries to solve, correctly detecting the negative samples (bad quality contours) would be more beneficial for the patients.

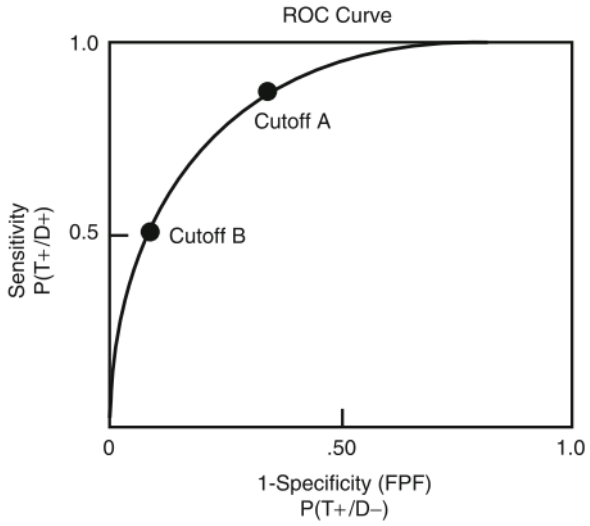


Figure 24: ROC curve (true positive rate versus false positive rate) showing two different cutoff points. Cutoff A has a high sensitivity but lower specificity, meaning that the amount of false positives is higher. Cutoff B has low sensitivity, and high specificity with the consequent higher amount of false negatives [66].

Therefore, the cut-off probability value for classifying the contours in the two classes needs to be modified from the default 0.5 value. As previously mentioned, choosing this cutoff point depends on the classification problem and its applications: it is necessary to evaluate if we care more about minimizing the false positives or the false negatives. For our research question, the clinical consequences of classifying an inaccurate contour as accurate (FP) is worse than classifying a good quality contour as a bad quality contour (FN). This is because the former option implies that a contour that is not good enough for plan selection will be incorrectly labeled as a good quality delineation, which would eventually lead to suggest a treatment plan from the plan library that is not the best fit for the anatomy-of-the-day. In the second case, a contour that is good enough for plan selection would be missed due to its incorrect labeling, potentially leading to longer plan selection time or the selection of a back-up plan. This does not have detrimental effects on the patient, but it would imply an increase in the treatment time and a higher workload for clinicians.

Consequently, the cutoff probability value should be set in a higher number than the default 0.5.

This would imply a worse ability of the classifier to detect accurate contours (lower sensitivity), but a better ability to classify inaccurate contours (higher specificity). However, this means that the number of false negatives is also increased. As previously mentioned, this approach has less clinical risk for the patient, but the trade-off between decreasing the false positives and increasing the workload for RTTs should be carefully evaluated when choosing the right cutoff point. Further discussion on how to perform a good selection of the cutoff value is done in section 5.6.

5.4. The importance of interpretability

In the last years, multiple studies have been published investigating different applications of AI in radiotherapy, including the automation of routine practices or improvement of certain workflows like OAR and tumor automatic segmentation, or automatic contour QA, which is the case of the work presented in this thesis [36].

In Chapter 2, multiple AI-based methodologies for automatic contour QA were explained showing how well the different ML and DL algorithms can perform when detecting bad quality delineations. However, it is necessary to take into account that the ultimate goal of all these AI-based strategies is their clinical implementation. To bring these algorithms into the clinic, accurate model performance is not the only factor to consider, their interpretability by medical physicists is also important, since they are the bridge between these automatic QA tools and their clinical implementation.

Interpretability is related to being able to understand the outcome or predictions of an algorithm, without the need of understanding the details of the algorithm's implementation or its mechanics [67]. When looking into the ML algorithms implemented in this project, logistic regression is considered "interpretable" contrary to the random forest classifier [67]. The parameters of logistic regression models are much easier to understand than from the random forest. The random forest has multiple parameters that allow tuning how the ensemble of decision trees classifies the input observations. Understanding how the decision trees and their consequent classification skills are affected when changing these parameters is much more difficult when the output is the result of majority voting.

However, it must be also acknowledged the trade-off between model performance and interpretability. There is no ML algorithm that has the highest accuracy and the highest interpretability [67]. Therefore, even though the RF classifier is more difficult to understand for medical physicists, it outperformed logistic regression, and for the clinical problem targeted in this project with the final goal of plan selection assessment, and not final decision-making, the RF model would be still the chosen methodology.

5.5. Limitations of the project

The lack of data availability leads to reduced data variance, which leads to introducing bias in the algorithms. Especially in ML, bias occurs when data from a certain class is more represented than the other class. For both, the bladder and the rectum, there is a difference between the amount of data available for each class: for the bladder, the bad quality contours are 42% of the dataset, while the accurate contours are the remaining 58%. However, the rectum dataset presents a more noticeable class imbalance that may have affected the results: 37% are inaccurate contours, while the remaining 63% are good quality delineations. Having fewer bad quality contours in the dataset makes this group less represented. Therefore, the ML algorithms have fewer chances to learn the ranges and distribution of feature data values that represent them.

Furthermore, this project is only focused on the QA of 3D contour volumes, which gives a limited insight to medical physicist about the contour quality. For a deeper and more detailed evaluation of the contours, it would be interesting to know the exact location of the contouring error in case the delineation is classified as a bad quality delineation. This would further aid medical physicists on quickly finding the problematic area and manually correct the contour. A possible strategy for implementing this is introduced in the next section.

5.6. Future work

Due to the time constraint for this project, there is still some improvements that can be implemented to obtain better results. First of all, it would be beneficial to improve the classification skills of the models by changing the cut-off probability from the default 0.5 value, to a value that fits better the clinical interests of this study. The suggestion would be to try different cutoff points, obtain the corresponding sensitivity,

specificity (from where false positives can be derived), and accuracy values. Then, choose the cutoff point with the best accuracy but considering as well the corresponding trade-off between sensitivity and specificity values. This should be done by cross validation on the training data and not on the test data to be sure that overfitting of the data is not happening when adjusting the threshold.

Moreover, as previously mentioned in section 5.5, the implemented methodologies analyze the feature data of the contour volumes, not slice by slice. An improvement would be to introduce a local error detection strategy to identify the exact slice and region in which the contour volume is not correct. For this purpose, it would be interesting to explore the potential of spatial probability maps (SPMs) as it was done by Van Rooij et al. in a study were they investigated the potential of SPMs as the local error detection strategy for DL-generated contours from the salivary glands [68].

For further automation of the QA process, DL is an option that could be explored as well. As explained in Chapter 2, DL has the potential of learning from the raw input data, without the tedious steps of feature extraction and selection which are necessary for machine learning implementation. The main limitation of implementing DL is that neural networks are complex learning algorithms that require a lot of data to properly train their parameters [36].

These suggestions for improving the methodology presented in this work have the final goal of developing an automatic contour QA tool that is more reliable and provides more insights into contour evaluation, with the local error detection strategy. Moreover, clinical validation would be another necessary step for future clinical implementation.

This thesis work can contribute to a full workflow for plan selection as an intermediate step. This workflow would start with an online auto-contouring tool on the CBCT scans, followed by the automatic contour QA tool developed in this thesis. The last step would be an algorithm acting as the bridge between the contour QA tool output and the suggested plan from the plan library. The goal of this algorithm would be to evaluate which treatment plan from the plan library matches better the AG contour's location and geometry. This would only happen if the AG contour is previously classified as a good quality delineation by the proposed automatic contour QA methodology.

6

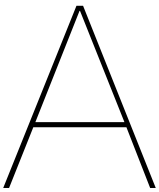
Conclusion

The quality of the OAR delineations obtained from online auto-contouring of CBCT scans from LACC patients is conditioned by their poor image quality. This study had the goal of designing a methodology capable of assessing the quality of these automatically-generated contours for evaluating their eligibility for plan selection assessment.

A comparison study between two ML-based algorithms (random forest and logistic regression) was performed. Due to the high impact of bladder and rectum filling on the tumor target position, automatically-generated contours from these OARs were used as the input datasets together with a few gold-standard contours available. These contours were manually labeled to have a binary classification problem (good vs. bad quality contours). Radiomic features providing a quantifiable measure of the phenotypic characteristics of the structures shown in the CBCT scans were obtained for each of the contours. RF and LR classifiers were used to infer statistical relationships from the training set between the OAR feature data and the class label. Predictions on the test data were obtained using the trained model to evaluate performance and classification skills. Random forest classifier gave the best results, especially for the bladder. However, it is not the preferred methodology when looking for high interpretability, but this is not a necessity for the intended application of this classifier. The contour QA methodology developed aims at providing a supporting tool for decision-making during plan selection. The final decision is made by the clinician, hence the work presented in this thesis is not the endpoint of the plan selection procedure. Having a classifier with more potential to correctly detect inaccurate contours is more important for the clinical application described. This allows trading-off interpretability for model performance and selecting the random forest as the preferred classifier for the automatic contour QA tool.

With the increasing automation of radiotherapy workflows, automatic procedures for QA are a necessity to ensure patient safety, while delivering the best possible treatment. The automatic contour QA tool developed would play a key role to ensure a faster, more feasible and consistent plan selection. Moreover, this thesis provides a starting point for standardizing automatic contour QA procedures with the goal of plan selection assessment, providing a methodology that has potential for LACC patients. This means that the strategy implemented could also have potential for other tumor sites with diverse inter-fraction motion. However, further work needs to be done for improving the specificity of the classifier and its robustness so that its clinical implementation is more feasible.

Appendices



PyRadiomics features

(See all the extracted features in the next page, in Table 5).

Table 5: All the 72 features extracted from the data using PyRadiomics python library

Feature Class	Feature Name	Feature Class	Feature Name
Shape	Elongation	**GLCM**	Autocorrelation
	Flatness		ClusterProminence
	LeastAxisLength		ClusterShade
	MajorAxisLength		ClusterTendency
	Maximum2DDiameterColumn		Contrast
	Maximum2DDiameterRow		Correlation
	Maximum2DDiameterSlice		DifferenceAverage
	Maximum3DDiameter		DifferenceEntropy
	MeshVolume		DifferenceVariance
	MinorAxisLength		Id
	Sphericity		Idm
	SurfaceArea		Idmn
	SurfaceVolumeRatio		Idn
	VoxelVolume		Imc1
First order	10Percentile		Imc2
	90Percentile		InverseVariance
	Energy		JointAverage
	Entropy		JointEnergy
	InterquartileRange		JointEntropy
	Kurtosis		MCC
	Maximum		MaximumProbability
	MeanAbsoluteDeviation		SumAverage
	Mean		SumEntropy
	Median		SumSquares
	Minimum	**GLRLM**	GrayLevelNonUniformity
	Range		GrayLevelNonUniformityNormalized
	RobustMeanAbsoluteDeviation		GrayLevelVariance
	RootMeanSquared		HighGrayLevelRunEmphasis
	Skewness		LongRunEmphasis
	TotalEnergy		LongRunHighGrayLevelEmphasis
	Uniformity		LongRunLowGrayLevelEmphasis
	Variance		LowGrayLevelRunEmphasis
			RunEntropy
			RunLengthNonUniformity
			RunLengthNonUniformityNormalized
			RunPercentage
			RunVariance
			ShortRunEmphasis
			ShortRunHighGrayLevelEmphasis
			ShortRunLowGrayLevelEmphasis

B

Correlation analysis data

B.1. Bladder

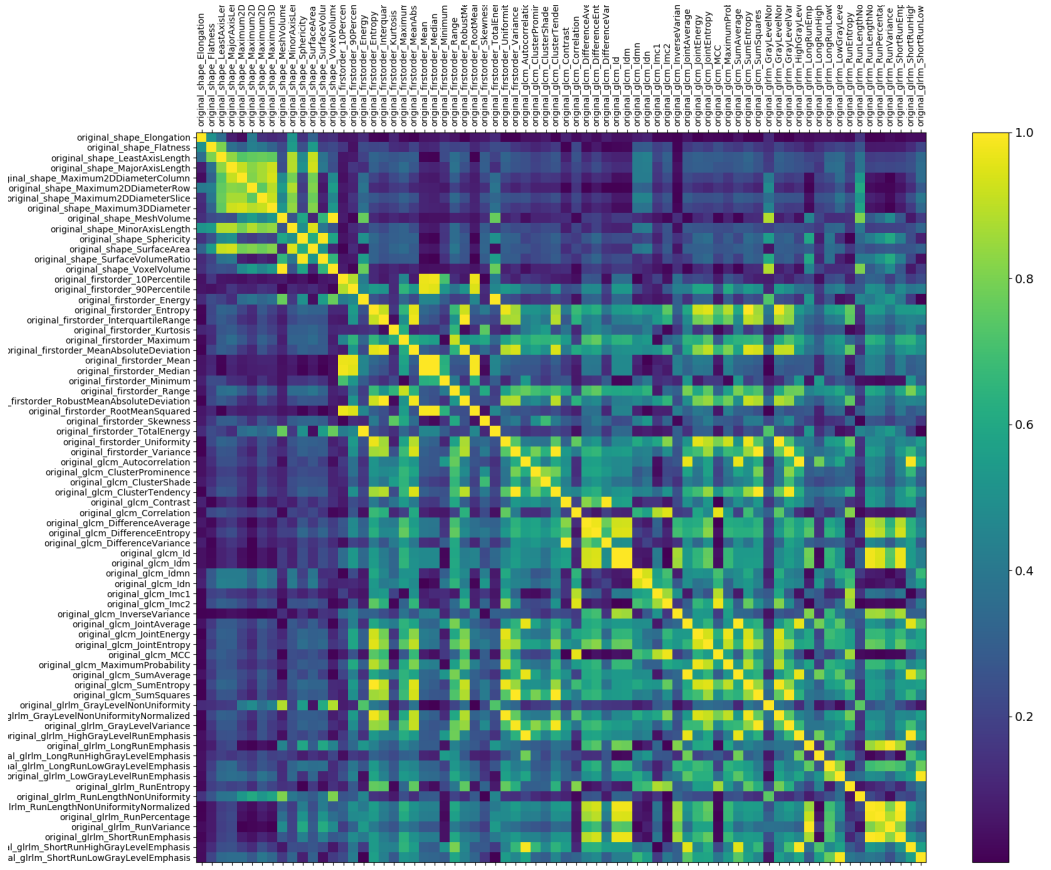


Figure 25: Correlation matrix for the 72 features extracted from the bladder. The strongly correlated pairs were identified by setting the condition of Pearson correlation coefficient > 0.8.

B.2. Rectum

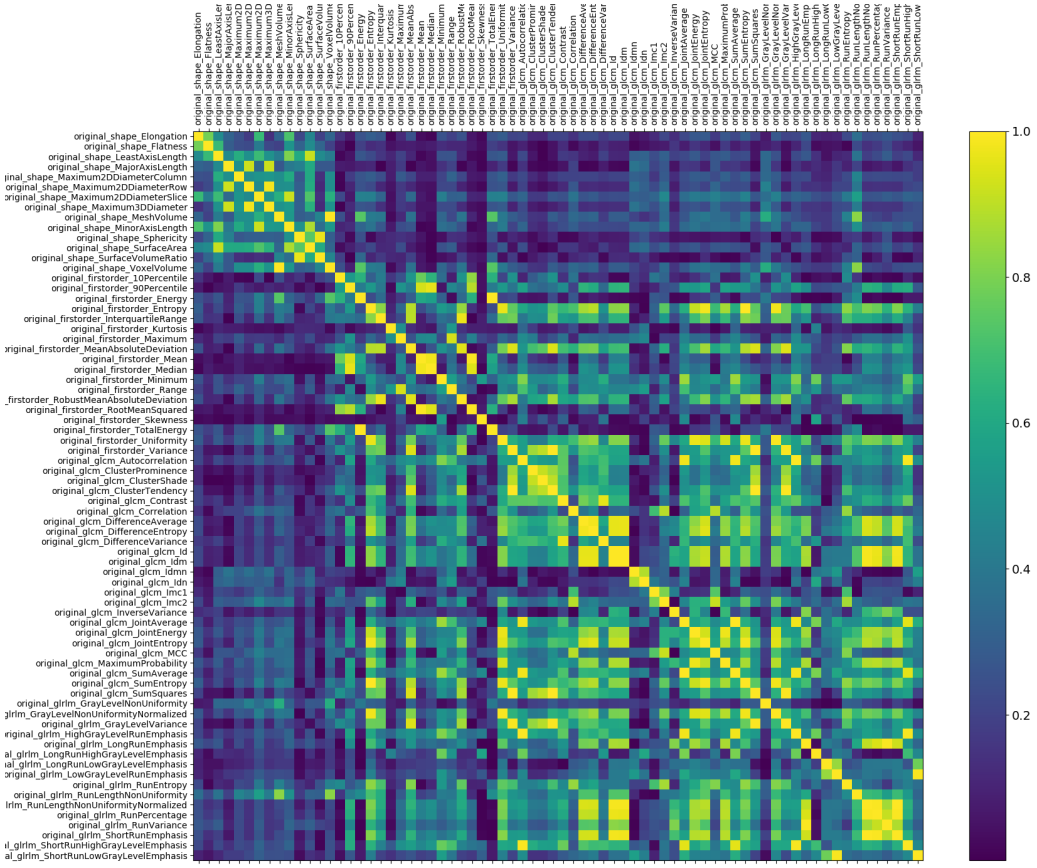


Figure 26: Correlation matrix for the 72 features extracted from the rectum. The strongly correlated pairs were identified by setting the condition of Pearson's correlation coefficient > 0.8.

Bibliography

[1] WHO. “Cervix uteri Source: Globocan 2020”. In: *International Agency for Research on Cancer (IARC)* 419 (2020), pp. 1–10. URL: https://gco.iarc.fr/today/data/factsheets/cancers/23-Cervix-uteri-fact-sheet.pdf.

[2] IKNL. *NKR-cijfers*. URL: https://iknl.nl/kankersoorten/baarmoederhalskanker.

[3] NVOG. *Cervixcarcinoom, landelijke richtlijn*. 2012. URL: https://richtlijnendatabase.nl/richtlijn/cervixcarcinoom/algemeen.html.

[4] Neerja Bhatla et al. “Revised FIGO staging for carcinoma of the cervix uteri”. In: *International Journal of Gynecology Obstetrics* 145 (2019), pp. 129–135.

[5] Kathrin Kirchheiner et al. “Health related quality of life and patient reported symptoms before and during definitive radio(chemo)therapy using image-guided adaptive brachytherapy for locally advanced cervical cancer and early recovery - A mono-institutional prospective study”. In: *Gynecologic Oncology* 136.3 (2015), pp. 415–423.

[6] Dominique M.W. Reijtenbagh et al. “Patient-reported acute GI symptoms in locally advanced cervical cancer patients correlate with rectal dose”. In: *Radiotherapy and Oncology* 148 (2020), pp. 38–43.

[7] Anouk Corbeau. *Clinical implementation of adaptive intensity-modulated proton therapy (aIMPT) for locally advanced cervical cancer (LACC)*. Tech. rep. Erasmus MC, 2021.

[8] *Cervical Cancer Treatment-Health Professional Version*. 2021. URL: https://www.cancer.gov/types/cervical/hp/cervical-treatment-pdq.

[9] Cyrus Chargari et al. “Brachytherapy: An overview for clinicians”. In: *CA: A Cancer Journal for Clinicians* 69.5 (2019), pp. 386–401.

[10] R. Jadon et al. “A Systematic Review of Organ Motion and Image-guided Strategies in External Beam Radiotherapy for Cervical Cancer”. In: *Clinical Oncology* 26.4 (2014), pp. 185–196.

[11] Richard Pötter et al. “The EMBRACE II study: The outcome and prospect of two decades of evolution within the GEC-ESTRO GYN working group and the EMBRACE studies”. In: *Clinical and Translational Radiation Oncology* 9 (2018), pp. 48–60.

[12] Jean Baptiste Guy et al. “Dosimetric study of volumetric arc modulation with RapidArc and intensity-modulated radiotherapy in patients with cervical cancer and comparison with 3-dimensional conformal technique for definitive radiotherapy in patients with cervical cancer”. In: *Medical Dosimetry* 41.1 (2016), pp. 9–14.

[13] Luca Cozzi et al. “A treatment planning study comparing volumetric arc modulation with RapidArc and fixed field IMRT for cervix uteri radiotherapy”. In: *Radiotherapy and Oncology* 89.2 (2008), pp. 180–191.

[14] Sabrina T. Heijkoop et al. “Quantification of intra-fraction changes during radiotherapy of cervical cancer assessed with pre- and post-fraction Cone Beam CT scans”. In: *Radiotherapy and Oncology* 117.3 (2015), pp. 536–541.

[15] André Buchali et al. “Impact of the filling status of the bladder and rectum on their integral dose distribution and the movement of the uterus in the treatment planning of gynaecological cancer”. In: *Radiotherapy and Oncology* 52 (1999), pp. 29–34.

[16] Rozilawati Ahmad et al. “Increasing treatment accuracy for cervical cancer patients using correlations between bladder-filling change and cervix-uterus displacements: Proof of principle”. In: *Radiotherapy and Oncology* 98.3 (2011), pp. 340–346.

[17] M. L. Bondar et al. “Individualized nonadaptive and online-adaptive intensity-modulated radiotherapy treatment strategies for cervical cancer patients based on pretreatment acquired variable bladder filling computed tomography scans”. In: *International Journal of Radiation Oncology Biology Physics* 83.5 (2012), pp. 1617–1623.

[18] Sabrina T. Heijkoop et al. “Clinical implementation of an online adaptive plan-of-the-day protocol for nonrigid motion management in locally advanced cervical cancer IMRT”. In: *International Journal of Radiation Oncology Biology Physics* 90.3 (2014), pp. 673–679.

[19] Sabrina T Heijkoop. “Plan-of-the-Day Adaptive Radiotherapy for Locally Advanced Cervical Cancer”. PhD thesis. Erasmus University Rotterdam, 2017.

[20] Cheukkai B. Hui et al. “Quality assurance tool for organ at risk delineation in radiation therapy using a parametric statistical approach”. In: *Medical Physics* 45.5 (2018), pp. 2089–2096.

[21] C. Robert et al. “Clinical implementation of deep-learning based auto-contouring tools–Experience of three French radiotherapy centers”. In: *Cancer/Radiotherapie* 25.6-7 (2021), pp. 607–616.

[22] E. Nováková et al. “What is the optimal number of library plans in ART for locally advanced cervical cancer?” In: *Radiotherapy and Oncology* 125.3 (2017), pp. 470–477.

[23] Liesbeth Vandewinckele et al. “Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance”. In: *Radiotherapy and Oncology* 153 (2020), pp. 55–66.

[24] Reza Kalantar et al. “Automatic Segmentation of Pelvic Cancers Using Deep Learning: State-of-the-Art Approaches and Challenges”. In: *Diagnostics* 11.11 (2021).

[25] Carlos E. Cardenas et al. “Generating High-Quality Lymph Node Clinical Target Volumes for Head and Neck Cancer Radiation Therapy Using a Fully Automated Deep Learning-Based Approach”. In: *International Journal of Radiation Oncology Biology Physics* 109.3 (2021), pp. 801–812.

[26] Abdel Aziz Taha and Allan Hanbury. “Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool”. In: *BMC Medical Imaging* 15.1 (2015).

[27] Tiezhi Zhang et al. “Automatic Delineation of On-Line Head-And-Neck Computed Tomography Images: Toward On-Line Adaptive Radiotherapy”. In: *International Journal of Radiation Oncology Biology Physics* 68.2 (2007), pp. 522–530.

[28] Dong Joo Rhee et al. “Automatic detection of contouring errors using convolutional neural networks”. In: *Medical Physics* 46.11 (2019), pp. 5086–5097.

[29] Hanne Nijhuis et al. “Investigating the potential of deep learning for patient-specific quality assurance of salivary gland contours using EORTC-1219-DAHANCA-29 clinical trial data”. In: *Acta Oncologica* 60.5 (2021), pp. 575–581.

[30] Amy Frederick et al. “A Framework for Clinical Validation of Automatic Contour Propagation: Standardizing Geometric and Dosimetric Evaluation”. In: *Practical Radiation Oncology* 9.6 (2019), pp. 448–455.

[31] R. Msika et al. “Evaluation of a software for automatic delineation of the mammary gland and organs at risk in patients treated for breast cancer in lateral position”. In: *Cancer/Radiotherapie* 24.8 (2020), pp. 799–804.

[32] Ying Song et al. “Automatic delineation of the clinical target volume and organs at risk by deep learning for rectal cancer postoperative radiotherapy”. In: *Radiotherapy and Oncology* 145 (2020), pp. 186–192.

[33] Elaine Cha et al. “Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy”. In: *Radiotherapy and Oncology* 159 (2021), pp. 1–7.

[34] Dong Joo Rhee et al. “Automatic contouring system for cervical cancer using convolutional neural networks”. In: *Medical Physics* 47.11 (2020), pp. 5648–5658.

[35] Gregory Sharp et al. “Vision 20/20: Perspectives on automated image segmentation for radiotherapy”. In: *Medical Physics* 41.5 (2014).

[36] Sunan Cui et al. “Introduction to machine and deep learning for medical physicists”. In: *Medical Physics* 47.5 (2020), e127–e147.

[37] Domingos Pedro. “A Few Useful Things to Know About Machine Learning”. In: *Communications of the ACM* 55.10 (2012), pp. 79–87.

[38] Alan M. Kalet, Samuel M.H. Luk, and Mark H. Phillips. “Radiation Therapy Quality Assurance Tasks and Tools: The Many Roles of Machine Learning”. In: *Medical Physics* 47.5 (2020), e168–e177. ISSN: 00942405.

[39] Chris McIntosh, Igor Svistoun, and Thomas G. Purdie. “Groupwise conditional random forests for automatic shape classification and contour quality assessment in radiotherapy planning”. In: *IEEE Transactions on Medical Imaging* 32.6 (2013), pp. 1043–1057.

[40] Hsin Chen Chen et al. “Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: A general strategy”. In: *Medical Physics* 42.2 (2015), pp. 1048–1059.

[41] M. B. Altman et al. “A framework for automated contour quality assurance in radiation therapy including adaptive techniques”. In: *Physics in Medicine and Biology* 60.13 (2015), pp. 5199–5209.

[42] J. Zhang, O. Ates, and A. Li. “Implementation of a Machine Learning–Based Automatic Contour Quality Assurance Tool for Online Adaptive Radiation Therapy of Prostate Cancer”. In: *Int J Radiat Oncol Biol Phys* 96.2 (2016), E668.

[43] David Polly. *Procrustes and PCA the details*. 2018. URL: `https://www.cnidaria.nat.uni-erlangen.de/shortcourse/Lecture%203%20-%20Procrustes%20and%20PCA.pdf`.

[44] Samsara Terparia et al. “Automatic evaluation of contours in radiotherapy planning utilising conformity indices and machine learning”. In: *Physics and Imaging in Radiation Oncology* 16.February (2020), pp. 149–155.

[45] Ying Zhang et al. “Texture-based, automatic contour validation for online adaptive replanning: A feasibility study on abdominal organs”. In: *Medical Physics* 46.9 (2019), pp. 4010–4020.

[46] Xinyuan Chen et al. “CNN-Based Quality Assurance for Automatic Segmentation of Breast Cancer in Radiotherapy”. In: *Frontiers in Oncology* 10.April (2020), pp. 1–9.

[47] Kuo Men et al. “Automated Quality Assurance of OAR Contouring for Lung Cancer Based on Segmentation With Deep Active Learning”. In: *Frontiers in Oncology* 10.July (2020), pp. 1–7.

[48] Janita E. van Timmeren et al. “Radiomics in medical imaging—“how-to” guide and critical reflection”. In: *Insights into Imaging* 11.1 (2020).

[49] Joost J.M. Van Griethuysen et al. “Computational radiomics system to decode the radiographic phenotype”. In: *Cancer Research* 77.21 (2017), e104–e107.

[50] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. “Radiomics: Images are more than pictures, they are data”. In: *Radiology* 278.2 (2016), pp. 563–577.

[51] Robert M. Haralick, Its’hak Dinstein, and K. Shanmugam. “Textural Features for Image Classification”. In: *IEEE Transactions on Systems, Man and Cybernetics* SMC-3.6 (1973), pp. 610–621.

[52] Sylvain Reuzé et al. “Radiomics in Nuclear Medicine Applied to Radiation Therapy: Methods, Pitfalls, and Challenges”. In: *International Journal of Radiation Oncology Biology Physics* 102.4 (2018), pp. 1117–1142.

[53] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[54] Leo Breiman. “Random Forests”. In: *Machine Learning* 45 (2001), pp. 5–32.

[55] Ian T. Jollife and Jorge Cadima. “Principal component analysis: A review and recent developments”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016).

[56] Kanish Shah et al. “A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification”. In: *Augmented Human Research* 5.1 (2020), pp. 1–16.

[57] David G. Kleinbaum and Mitchel Klein. *Logistic Regression: A Self-Learning Text*. 3rd. Springer, 2010.

[58] Saed Sayad. *Logistic Regression*. URL: https://www.saedsayad.com/logistic_regression.htm.

[59] Miranda E.M.C. Christianen et al. “Predictive modelling for swallowing dysfunction after primary (chemo)radiation: Results of a prospective observational study”. In: *Radiotherapy and Oncology* 105.1 (2012), pp. 107–114.

[60] Issam El Naqa et al. “Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors”. In: *International Journal of Radiation Oncology Biology Physics* 64.4 (2006), pp. 1275–1286.

[61] Isabelle Guyon and André Elisseeff. “An Introduction to Variable and Feature Selection”. In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[62] Jason Brownlee. *Tour of Evaluation Metrics for Imbalanced Classification*. 2021. URL: https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/.

[63] *Classification: ROC Curve and AUC*. 2020. URL: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

[64] Miguel Ángel González Ballester, Andrew P. Zisserman, and Michael Brady. “Estimation of the partial volume effect in MRI”. In: *Medical Image Analysis* 6.4 (2002), pp. 389–405.

[65] Raymond Miralbell et al. “Radiotherapy of bladder cancer: Relevance of bladder volume changes in planning boost treatment”. In: *International Journal of Radiation Oncology Biology Physics* 41.4 (1998), pp. 741–746.

[66] Sylvia Wassertheil-Smoller and Jordan Smoller. “Cutoff Point and Its Effects on Sensitivity and Specificity”. In: *Biostatistics and Epidemiology: A Primer for Health and Biomedical Professionals*. 4th. 2015. Chap. 5, pp. 140–142.

[67] Yi Luo et al. “Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling”. In: *BJR|Open* 1.20190021 (2019).

[68] Ward van Rooij et al. “Using Spatial Probability Maps to Highlight Potential Inaccuracies in Deep Learning-Based Contours: Facilitating Online Adaptive Radiation Therapy”. In: *Advances in Radiation Oncology* 6.2 (2021), p. 100658.