

MENG INDIVIDUAL PROJECT

DEPARTMENT OF COMPUTING

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

**[INTERIM]: Transfer Learning for Deep Learning
Radiotherapy Planning**

Author:
Anton Zhitomirsky

Supervisor:
Ben Glocker

Second Marker:
TODO

January 21, 2024

Abstract

 First draft of abstract

Clinicians target cancerous tumours by studying 3D contrasting images of cancerous tumours and surrounding soft tissues to plan targets for radiation therapy. The Royal Marsden Hospital is a key contributor of data for this project, which uses this approach to delineate tumours for cervical cancers. Typically after a gross tumour volume (GTV) is extrapolated from the relevant imaging modality, clinicians append tailored safety margins to also account for the microscopic cancerous spreads not visible in the scan to generate the planned target volume (PTV).

The PTV area has to be generous enough to attempt to treat the problem in one-shot, yet conservative enough to not harm surrounding healthy tissue with radiation over the course of the treatment. Compounded with small sample size of labelled data this proposes a significant challenge for developing deep-learning segmentation models to identify an optimal PTV.

Thus we propose a transfer learning strategy to utilize imaging models in similar domains to attempt to learn from the limited input size to provide clinicians with a faster and more accurate segmentation method.

Contents

1	Introduction	2
1.1	Clinical Context	2
1.2	Motivation	2
1.3	Current Solutions	2
1.4	Outline of Report	2
2	Background	3
2.1	Clinical Context	3
2.2	Vanilla Image Segmentation Models	3
2.2.1	Convolutional Neural Networks (CNN)	3
2.2.2	U-Net	3
2.2.3	nnU-Net	3
2.3	Existing Auto-Segmentation Methods	3
2.3.1	Total Segmentator	3
2.3.2	UniverSeg	3
2.3.3	SAM	3
2.4	Current Limitations	3
2.4.1	Data Size	3
2.4.2	Bespoke Application	4
2.5	Transfer Learning	4
3	Data	5
3.1	CT scan	5
3.2	File Format	5
3.3	Notes	6
3.4	Delineation classes	6
3.4.1	Organs At Risk	7
3.4.2	CTVp	7
3.4.3	CTVn	8
3.4.4	Parametrium	8
3.5	Establishing Rules for Structures	9
3.5.1	Relationship between Structures	9
3.5.2	Rules	9
3.6	Data Pre-Processing	10

4	Evaluation Metrics	11
4.1	Classification Based	11
4.2	Spatial Overlap Based	12
4.3	Surface Based	12
4.4	Volume Based	12
4.5	Evaluation	13
4.6	Estimated Editing Based	13
4.6.1	Surface DSC	13
4.6.2	Added Path Length	14
4.7	Summary	14
5	Proposal	15
5.1	Baseline Results	15
5.1.1	nnU-Net	15
5.1.2	Total Segmentator	15
5.1.3	UniverSeg	15
5.2	Summary	15
6	Interim Deliverables	16
6.1	Project Plan	16
6.2	Evaluation Plan	16
7	Ethics	17
7.1	Patient disclosures	17
7.2	Using the tool	17
	Bibliography	20

Chapter 1


Introduction

1.1 Clinical Context

1.2 Motivation

1.3 Current Solutions

1.4 Outline of Report

 This interim report will not contain implementation details yet.
--

The structure of this report will take the reader of a reasonable scientific background through the project such that they might be able to reconstruct the outcome themselves. It is expected of the reader to understand the concept of Machine Learning and an intuition surrounding Computer Vision. Firstly Chapter 2 discusses the clinical context (Section 2.1) from which the idea for this project originates. We further construct the background knowledge for the current Computer Vision methodologies which exist to solve segmentation based issues (Section 2.2). This will lead into currently pre-trained models (Section 2.3) and their application in a Transfer Learning setting (Section 2.5).

With collaboration from the Royal Marsden Hospital they have provided this project with a set of data (Chapter 3) with which the remainder of this project is trained upon, the usage concerns have been evaluated through the Ethics chapter (Chapter 7). We discuss what constitutes an accurate segmentation and reason about evaluation metrics to provide an outlook on the models performance (Chapter 4).

Finally, Chapter 5 contains details surrounding the proposition of this project. We establish base-line results (Section 5.1) such that we may see that our future implementation will supersede the common implementation. Currently, for the interim report the proposition lays bare, but will be slowly added throughout the course of the project.

Chapter 2

Background

2.1 Clinical Context

2.2 Vanilla Image Segmentation Models

2.2.1 Convolutional Neural Networks (CNN)

2.2.2 U-Net

2.2.3 nnU-Net

2.3 Existing Auto-Segmentation Methods

2.3.1 Total Segmentator

2.3.2 UniverSeg

2.3.3 SAM

2.4 Current Limitations

The approaches listed in Section 2.2 and models discussed in Section 2.3 are great approaches for most image segmentation applications. We've seen the advancements of CNNs to tackle the intractable nature of fully connected neural network, and the advancements in segmentation models in the U-Net. Furthermore, these techniques have been used to train robust models in medicine such as those presented in Section 2.3.

However, our problem is not accurately solved with the methods mentioned. This is due to a handful of independent details which require more careful planning and engineering.

2.4.1 Data Size

The data quantity supplied is a limiting factor for creating a robust model. We are given 100 labeled data elements across 5 classes. Without vast collection of knowledge, it is hard for an application to create a model which generalizes well to the total population, especially in a very specific and bespoke use case as radiotherapy planning for cervical cancer.

2.4.2 Bespoke Application

Another issue lies in the bespoke nature of this application. Most pre-trained networks currently run segmentation on structures that are more obvious in a given image modality. For instance, TotalSegmentator has learnt a robust model for delineating 117 classes of objects in the human body, such as bones, a large subset of significant organs and veins [25].

2.5 Transfer Learning

Transfer Learning uses knowledge that has been obtained from one task, and uses it as a starting point for learning a new task. It is therefore a useful solution to the problems identified in Section 2.4 because of the transferable knowledge features for similar domains and its proven success in generalizing features is trained properly.

The intuitive reason why transfer learning works is because in the early layers of deep learning, the model learns very low-level features. At this scale, the initial data-set or the cost function doesn't matter because a model working on the same problem but with different initialization will learn similar low-level features. This allows transferral because the (large/sufficiently sized) input dataset is abstracted in the set low-level features which can instead be transferred. Then, the later layers are more specialized to a particular task [3]. It is similar to seeing the distribution in the training data change and transferring knowledge across domains [17].

Transfer Learning has the potential to: improve initial performance using only the transferred knowledge before any further learning is done, improve the time it takes to fully learn the target task given the transferred knowledge, and improve the final performance all when compared to initial benchmarks without transfer [22]. It has also been found to work in medical contexts as well, where, for 332 abdominal liver CT scans, transfer learning generally improved weight initialization and resulted in faster convergence providing stronger and more robust representation [9].

Transfer Learning has been seen to prevent overfitting in domains where data volume is low and where generality without overfitting is hard to come by. This is because the model has already learnt features that are likely to be useful in the second task [26].

However, generalization is not a guarantee, as overfitting is still possible if the model is fine-tuned too much on the second task, as it may 'learn task-specific features that do not generalize well to new data' [26]. In our case, our target dataset is small, but similar to the base network dataset. Here, we may overfit because fine-tune the pre-trained network with the target dataset may not generalize to the global population. If instead we attempt to transfer a task with different base network dataset, then using high-level features of the pre-trained model will not be useful [26].

Chapter 3

Data

The data is acquired during a CT scan (Section 3.1) and presented as a set of NIfTI (Section 3.2) files provided by the Royal Marsden Hospital. The data is of 100 patients each with a variant of cervical cancer. We have obtained from the hospital a spreadsheet with additional notes about each patient which may be useful in training and debugging (Section 3.3). Finally, this data is labelled into 5 different classes as a binary segmentation problem (Section 3.4). Included is a set of 10 hold-out data items, which are patients with only the raw CT scan with no labels.

3.1 CT scan

Before we consider other aspects of the data it is helpful to consider the context from which it was extracted and therefore what we might expect to see. This data is in CT scan, and so will be the focus, although there exist other imaging modalities such as Magnetic Resonance Imaging (MRI) and others.

A CT Scan is an X-ray study, where a series of rays are rotated around a specified body part, and computer-generated cross-sectional images are produced [10]. The granularity or image slice thickness is decided by the operator or physician and ranges from 1mm to 10mm. Whilst the scanner rotates the X-ray tube the patient is slowly moved up or down in the table to produce different cross-section images.

We therefore expect to receive a representation of the internal structure or functions of an atomic region in the form of an array of voxels. A voxel represents the value on a grid in three-dimensional space and is decided by the physician once they establish the slice thickness.

3.2 File Format

The files are stored in a .nii file format which defines a style of image called the ‘Neuroimaging Informatics Technology Initiative’ (NIfTI) [10]. It serves as a lightweight alternative to other formats such as DICOM and eliminates ambiguity from spatial orientation information [12].

The file has a fixed-size header which stores information about the data collected. Table 3.1 summarizes some key attributes of the header. All other attributes not listed are handled by the SimpleITK library [2] which we use to read and manipulate the data in this project. The library defines the image as a set of points in a grid occupying a physical region in space as defined by this metadata, and therefore is influenced by the origin, size, spacing and so on.

Name	Meaning	Value example
dim	Image dimension	3 512 512 193 1 1 1 1
bitpix	Number of bits per voxel	32
pixdim	The grid spacing (voxel size) and optionally time interval	0 1.3 1.3 2.5 0 0 0 0
xyzt_units	indicates units of pixdim and defined in the C header, e.g. NIFTI_UNITS_MM = 2	2

Table 3.1: Description of NIfTI header parameters relevant to this project [12, 7, 4]. Example values are taken from patient id:075.

3.3 Notes

The notes contain information about each of the 100 labelled data pairs [5]. This information can be helpful in debugging or troubleshooting. It also provides a good warning regarding the variability of the data. In particular some aspects to note are summarized in Table 3.2.

Patient ID	Comment	Concern
zzAMLART003	“no GTVp”	Some scans contain no visible tumour, but we still draw a CTV
zzAMLART017	“only scanned bottom of kidneys”	We should be cautious of variability of width scope given in our source data
zzAMLART017	“missing left kidney”	Unusual body anatomy might trip up the model, mentioned elsewhere are also
zzAMLART041	“extra slices”	variability in voxels or quantity may require data pre-processing to eliminate data uncertainty
zzAMLART055	“no contrast - hard to see LNs. NG tube in situ. posterior renal vein. small parametrium and low uterus”	An edge case like this will require more thought

Table 3.2: A few captivating notes about each patient and why it might be concerning

There exist note entires for the majority of the 100 patients which weren’t shown in Table 3.2. Regardless, these notes are helpful to identify what type of pre-processing we must do in order to fully address some differences between patients. The concern is to not overfit on the ‘normal’ cases but also generalize and engineer a solution that is also open-minded to extreme or poorly captured cases; there is a vast variability in the anatomy of patients which makes computer vision tasks more challenging.

3.4 Delineation classes

The clinicians at the Royal Marsden Hospital have provided segmentation labels for 5 high-priority classes of interest. These are the Bladder, Anorectum, CTVn, CTVp, and Parametrium.

3.4.1 Organs At Risk

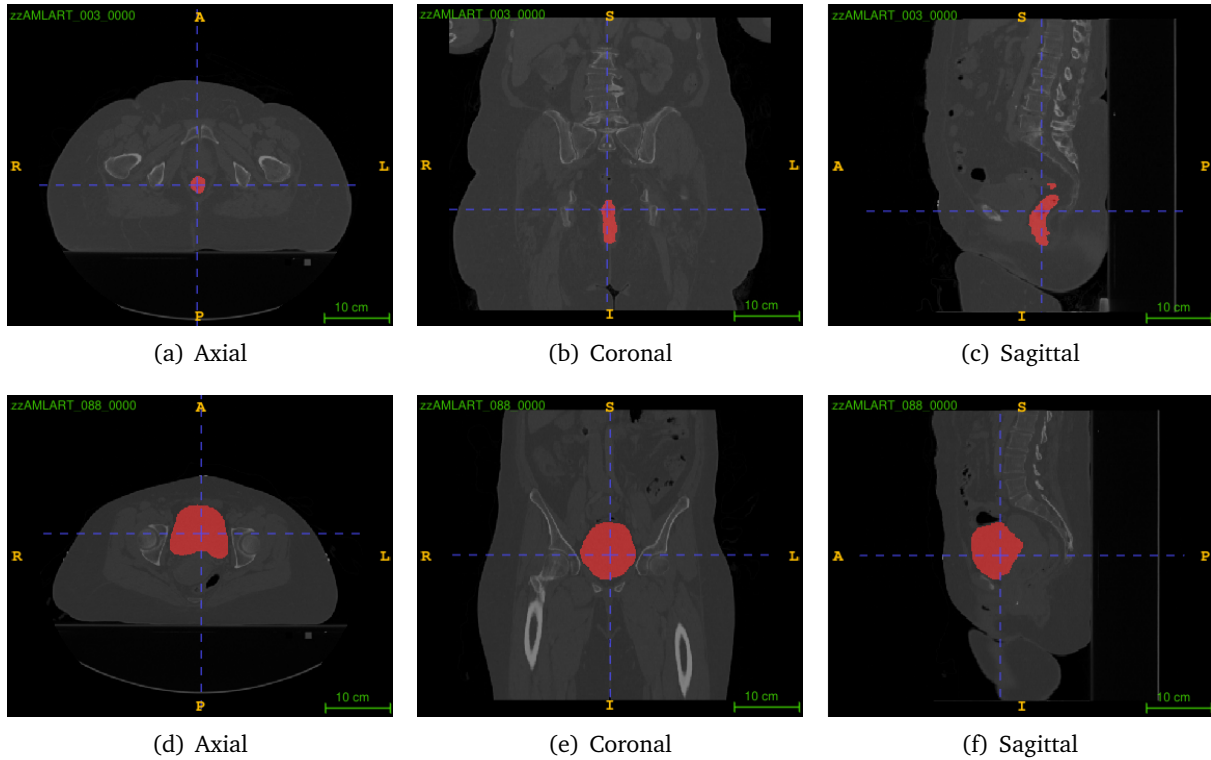


Figure 3.1: Views of a segmented (in red) Anorectum (3.1(a)-3.1(c)) and Bladder (3.1(d)-3.1(f)) of an arbitrary patient

An organ at risk is an organ which has a substantial probability of being within the PTV despite being healthy. Any areas that are created around the area should actively avoid these organs because by overlapping with them we risk complicating the treatment and compromising the health of functioning organs.

Many anatomies have been provided in the risk categories, however, in particular we have been supplied with contours for the Bladder (Figure 3.1(d)-3.1(f)) and the Anorectum (Figure 3.1(a)-3.1(c)). In particular, clinicians have identified that the Bladder may overlap with the CTVn (Section 3.4.3) and the Parametrium (Section 3.4.4).

3.4.2 CTVp

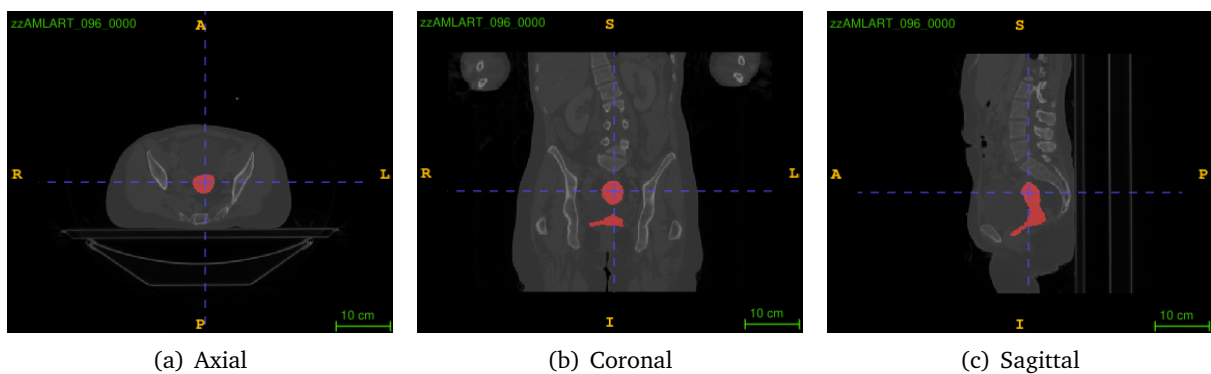


Figure 3.2: Views of a segmented (in red) CTVp of an arbitrary patient

The CTVp stands for the Primary Clinical Target Volume, see the example at Figure 3.2. This is the CTV where there may be local microscopic spread (uterus, cervix, upper vagina, primary tumour) [5]. This is the area that contains the tumour.

This isn't by any means an organ in a body, but rather an area comprised of other components formed by joining other structures together. The CTVp is an area defined in Equation 3.4.

3.4.3 CTVn

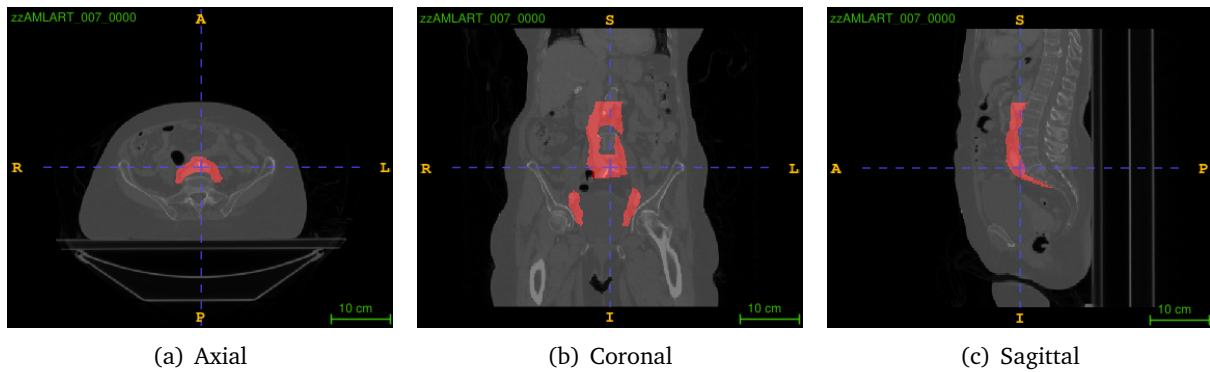


Figure 3.3: Views of a segmented (in red) CTVn of an arbitrary patient

The CTVn stands for Nodal Clinical Target Volume, see the example at Figure 3.3. This is the CTV where there may be microscopic spread to lymph nodes. It is drawn based on set margins around pelvic blood vessels and includes pelvic lymph nodes, common iliac lymph nodes and para-aortic lymph nodes [5].

Similarly to CTVp, this is a compound area with three groups of lymph nodes. In clinical practice, the number of these groups included in the CTV varies in each patient, depending on how advanced the disease is. Pathological lymph nodes (GTVn) are also included. The CTVn is an area defined in Equation 3.3.

3.4.4 Parametrium

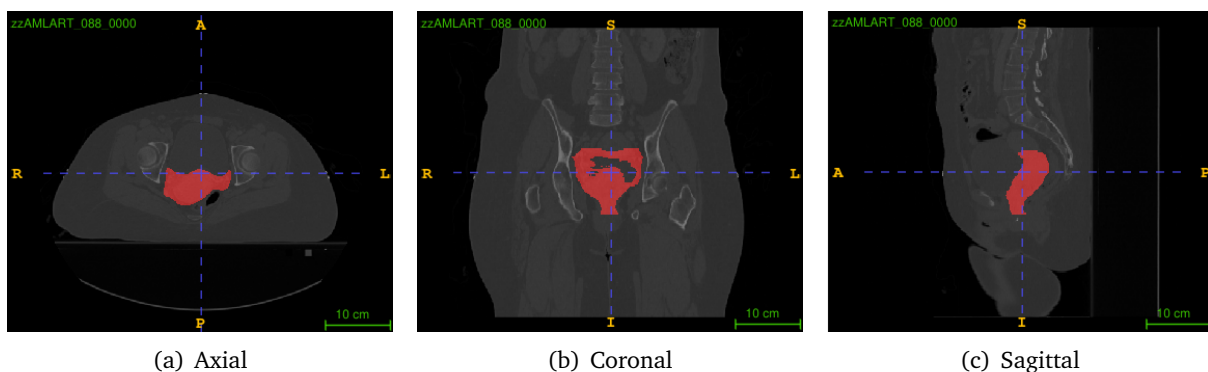


Figure 3.4: Views of a segmented (in red) Parametrium of an arbitrary patient

The Parametrium (or Paravagina) is the tissue surrounding the cervix/vagina - at risk of local spread, see Figure 3.4. Drawn as a complete structure and editing back to the level of vagina to be included [5].

3.5 Establishing Rules for Structures

Notation of Structures

- | | |
|---|---|
| 1. Let the Anorectum be denoted as A | 11. Let the Parametrium be denoted with P |
| 2. Let the Bladder be denoted as B | 12. Let the Uterus be denoted with U |
| 3. Let the Cervix be denoted with C | 13. Let the Vagina be denoted with V |
| 4. Let the CTVn be denoted with C_n | |
| 5. Let the CTVp be denoted with C_p | |
| 6. Let the GTVp be denoted with G_p | |
| 7. Let the GTVn be denoted with G_n | |
| 8. Let the Pelvic Lymph Node be denoted as L_p | |
| 9. Let the Common Iliac Lymph Node be denoted as L_i | |
| 10. Let the Para-aortic Lymph Node be denoted as L_{pa} | |

3.5.1 Relationship between Structures

- Let O denote the set $O = \{B, A, C_n, C_p, P\}$ for a particular patient. If we want to talk about a specific patient, we should use the super-script notation to differentiate patients, e.g., $O^i = \{B^i, A^i, C_n^i, C_p^i, P^i\}$.
- Let the overlap of two structures be denoted by the set intersect symbol \cap .
- Let the joint area of two structures be denoted by the set union symbol \cup .

3.5.2 Rules

The top 5 priority structures have been selected to identify and plan an area where radiotherapy should be used. With these structures, there are rules that the clinicians have outlined, they are quoted for clarification (these structures only refer to each independent patient):

- There should be no overlap between the CTVn, CTVp or Anorectum.

$$\forall i, j \in \{C_n, C_p, A\} \text{ with } i \neq j, i \cap j = \emptyset \quad (3.1)$$

- The Parametrium may overlap with all of the other structures.

$$\forall i \in S, \quad P \cap S_i \neq \emptyset \quad (\text{Possibly}) \quad (3.2)$$

- The Bladder may overlap with the CTVn.

$$B \cap C_n \neq \emptyset \vee B \cap C_n = \emptyset \quad (3.3)$$

- The CTVp is defined as a compound structure containing:

$$C_p = \overbrace{C \cup G_p}^{\text{High Risk CTV}} \cup U \cup V \quad (3.4)$$

- The CTVn is defined as a compound structure containing:

$$C_n = G_n \cup L_i \cup L_p + L_{pa} \quad (3.5)$$

3.6 Data Pre-Processing

👉 TODO

Chapter 4

Evaluation Metrics

To determine if a contour can be used in a clinical context, would be include calculating the difference between the provided labelled data. However, in a delineation context, we have different ways to evaluate this measure.

Suppose we are writing a linear-regression model to match a line onto a set of points. To quantify the performance of our line we would measure the shortest distance between each point and the predicted line. This relies on the assumption of points in a known domain that a model is attempting to approximate. In this case we are fitting a 1-dimensional model onto 0-dimensional points in the grid space.

However, it is far harder to decide on a scoring system when in a delineation context. Consider a single slice of a CT-scan with a known contour around the perimeter of a tumour¹. A model like those mentioned in Section 2.2 would attempt to learn a function to closely replicate the contour. Here our domain, prediction and ground truth are all 2-dimensional objects.

Geometric measures are the most popular, a survey has found [14]. These measures compare an auto-contour to a ground-truth contour and return a relative score based on its performance.

4.1 Classification Based

Assesses if voxels within and outside the auto-contour have been correctly labelled [14]. To begin, we define 'positive' to mean that the voxel selected is indeed in need of radiotherapy treatment, and 'negative' to mean that the voxel is classified as healthy.

A standard measure of classification is accuracy. It measures the total amount of correct predictions vs the total predictions it made. However, this measure alone isn't enough to fully capture the bias of a model because it doesn't tell you the full story with class-imbalanced data when there isn't an even number between positive and negative labels.

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

Better measures are Precision and Recall scores. The Precision (also known as the Positive Predictive Value [21]) measures the proportion of predictions that were successfully correct. The Recall (also known as True Positive Rate [21]), on the other hand, "measures the portion of positive voxels in the ground truth that are also identified as positive by the segmentation being evaluated".

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

¹Here we assume that the contour will hug the GTV tightly with no concern for microscopic spread around the remainder of the system

4.2 Spatial Overlap Based

Similarly to Classification Based metrics in Section 4.1, an Overlap Based metric measures the extent of overlap between an auto-contour and a reference structure [14].

The scores above can be combined into a more general score F_β to give

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

A specific case of this equation with $\beta = 1$ is mathematically equivalent to the DICE Similarity Coefficient which was found to be the most popular evaluation metric amongst 2021 studies [14, 21, 18].

$$F_1 = \text{DICE} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2TP + FP + FN} = \frac{2|S_g \cap S_p|}{|S_g| + |S_p|}$$

Where S_g is the ground truth segmentation and S_p is the predicted segmentation. From this relationship, the DICE score has found popularity in image segmentation for similar reasons that the F_1 score has found its popularity classical machine learning; it is able to provide a fair result for imbalanced datasets. This mentality is applicable in our scenario because a tumour will make up very little of the total volume of the domain space. This can be extended to a Volumetric DSC by considering the above in all 3-dimensions [24].

Another popular related evaluation method is the Jaccard Index, which measures the intersection over the union of two sets:

$$\text{JAC} = \frac{TP}{TP + FP + FN} = \frac{|S_g \cap S_p|}{|S_g \cup S_p|} \iff \frac{\text{DICE}}{2 - \text{DICE}}$$

Since the numerator for the Jaccard Index is smaller (since we avoid the issue of counting the intersecting sections twice) the JAC is always larger than the DICE score.

4.3 Surface Based

Also commonly known as Boundary-Distance-Based Methods [27] compares the distance between two structure surfaces. These can be either maximum distance, average distance or distance at a set percentile of ordered distances [21].

A common example is the Hausdorff Distance. Here, a directed distance metric is defined as the maximum distance from a point in the first set to a nearest point in the other between two individual voxels [27]. Therefore, the better the HD metric, the smaller the value it returns. Here, the distance is taken by some norm, typically Euclidian distance.

$$\text{HD}(A, B) = \max(h(A, B), h(B, A)), \quad \text{and directed } h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\|$$

The HD is generally sensitive to outliers. Because noise and outliers are common in medical segmentations, it is not recommended to use the HD directly [27]. Therefore, we can calculate the average directed Hausdorff Distance.

4.4 Volume Based

Volume-based metrics consider only the volume of the segmentation [15, 14, 27]. However, due to its poor spatial descriptions it is more commonly used jointly with other metrics.

$$\text{Relative Volume Difference (RVD)} = \left| \frac{|S_g| - |S_p|}{|S_g|} \right|$$

4.5 Evaluation

All these methods can be advantageous in some places rather than other. We can begin to list off some challenging scenarios to decide which segmentation is the best.

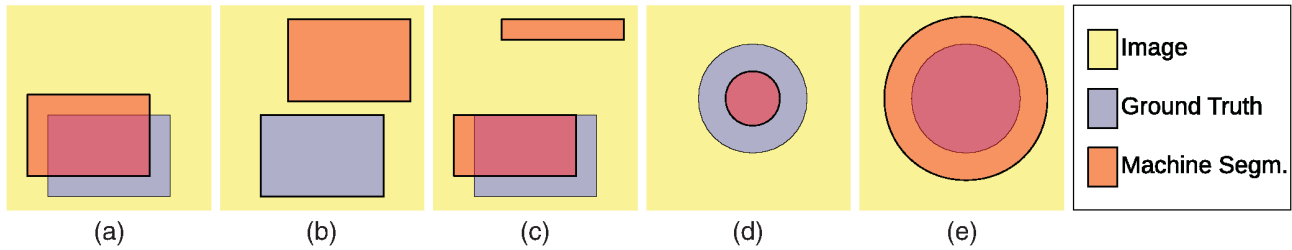


Figure 4.1: Figure from [27] illustrating cases of segmentation to aid with explanation of set-backs of certain evaluation metrics

- Classification Based (Section 4.1) and Spatial Overlap Based (Section 4.2) are similar; they are concerned with the number of correctly classified or misclassified voxels without taking into account their spatial distribution. Here, Figure 4.1(a) and Figure 4.1(c) would achieve similar results despite Figure 4.1(a) being locally bound to a better area.
- With Hausdorff Distance (Section 4.3) output segmentations generated by Figure 4.1(d) and Figure 4.1(e) will result in the same score, which is not favorable in a radiotherapy planning environment where an organ-at-risk is involved.
- Figure 4.1(b) would score flawlessly when using volumetric score estimation, however, it doesn't take into account spatial placement, which makes this measurement rather poor when used individually.

4.6 Estimated Editing Based

☞ This is a quotation from this paper, [18], however, it is referencing a paper of its own. Shall I reference the original paper or are 'linked' references OK?

It is difficult to select a measurement which can reflect a clinicians acceptability score. A study found that there was a lack of correlation between a geometric index and expert evaluation, with the JAC score having a 13% False Positive Rate. The conclusion of the study summarised that scores such as JSC and volumetric DSC, “provide limited clinical context and correlation with clinical or dosimetric quality” [18].

Because of the clinical context of evaluating the segmentation by a machine, it may sometimes be helpful to define a performance metric as the “fraction of the surface that needs to be redrawn” [16] since models at this point require manual review to avoid automation bias (Section 7.2). For larger structures, this method is useful it doesn't assign a lot of weight on the large trivial internal volume which accounts for a much larger proportion of the score.

4.6.1 Surface DSC

The study at [18] helped drive an initiative to combine aspects of Surface Based evaluation (Section 4.3) and Spatial Overlap Based evaluation (Section 4.2) into a Surface DICE. This assesses the

specified tolerance instead of the overlap of the two volumes.

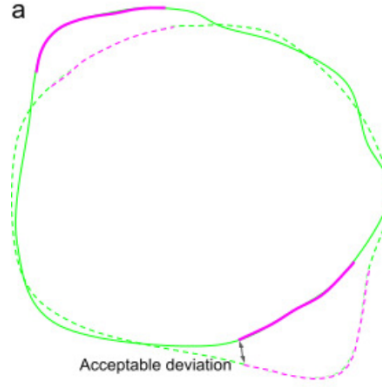


Figure 4.2: Taken from [16]. Illustrates the computation of the surface DICE, where the continuous line is the predicted surface and the dashed line is the ground truth. The black arrows show the maximum deviation tolerated without penalty; therefore, in pink is the unacceptable deviations and green otherwise.

We can formulate the Surface DSC score in a mathematical definition [18].

$$\text{Surface DSC} = \frac{|S_p \cap B_{g,\tau}| + |S_g \cap B_{p,\tau}|}{|S_p| + |S_g|}$$

Which provides a measure of the agreement between just the surfaces of two structures above a clinically determined tolerance parameter, τ . Here, $B_{p,\tau}$ represents the boundary region of the predicted surface within a maximum margin of deviation τ and similarly for $B_{g,\tau}$ for the ground truth.

4.6.2 Added Path Length

In a similar spirit, the APL was proposed as a score to predict “the path length of a contour that has to be added” [24]. This is achieved similarly by considering the number of added voxels required between the prediction and the gold standard with no regard to tolerance as a pose to Surface DSC (Section 4.6.1)

🔗 For future reference, *stack overflow discussion*
Implementation of surface DSC and APL: *source code*

4.7 Summary

This is why we settle at the Surface DSC (Section 4.6) which prioritizes deviation along boundary to a certain degree while measuring the fraction of the surface that needs to be redrawn, thus favouring a more conservative prediction of Figure 4.1(d) instead of (e).

For the purpose of this project, we shall select a evaluation measurement which is more bias towards conservative boundary estimates to not touch the organs at risk. This choice was in-part influenced by the clinician’s review pipeline; it would easier to correct Figure 4.1(d) instead of (e) because correcting the latter would likely take a considerable amount of time as it would require redrawing almost all of the boundary, whereas the former could be corrected much faster [16].

Chapter 5

Proposal

5.1 Baseline Results

👉 TODO for all

5.1.1 nnU-Net

5.1.2 Total Segmentator

5.1.3 UinverSeg

5.2 Summary

👉 Should we keep labels individual and segment them separately or should we segment it all in one shot?

👉 Solution to resolution problem, down-sample all samples or think of average?

Chapter 6

Interim Deliverables

6.1 Project Plan

You should explain what needs to be done in order to complete the project and roughly what you expect the timetable to be. Don't forget to include the project write-up (the final report), as this is a major part of the exercise. It's important to identify key milestones and also fall-back positions, in case you run out of time. You should also identify what extensions could be added if time permits. The plan should be complete and should include those parts that you have already addressed (make it clear how far you have progressed at the time of writing). This material will not appear in the final report.

6.2 Evaluation Plan

Project evaluation is very important, so it's important to think now about how you plan to measure success. For example, what functionality do you need to demonstrate? What experiments do you need to undertake and what outcome(s) would constitute success? What benchmarks should you use? How has your project extended the state of the art? How do you measure qualitative aspects, such as ease of use? These are the sort of questions that your project evaluation should address; this section should outline your plan.

Chapter 7

Ethics

This project involves very intimate and personal information of many female patients. Researchers may collaborate with third-parties by providing anonymized data which may not be reverse engineered back to the patient. The lack of this effort may result in “stigma, embarrassment, and discrimination” [19] if the data is misused.

7.1 Patient disclosures

The Royal Marsden Hospital doesn’t require “explicit consent” for sharing collected clinical data with outside entities as long as the patient is made aware of the ways their “de-identified/anonymized” data may be used [23]. Formalities are also arranged with Imperial Collage’s Medical Imaging team such as acting as “ethical data stewards” [11]. Without such disclosure and anonymisation of data, patients may be reluctant to provide candid and complete disclosures of their sensitive information, even to physicians, which may prevent a full diagnosis if their data isn’t maintained in an anonymous fashion.

The MIRA team acts as responsible data stewards by storing anonymized data within a folder on the college network. All provided data was anonymized by the Royal Marsden Hospital and sent to team MIRA in the NIfTI file format which discloses no personal identifiable information, as defined by GOV website [8]. This folder contains security measures which limit the availability of data only to those with specific access rights. Furthermore, operating on the preamble of de-identified data further reduces individual patient risk in the event that data is ever brought outside the confines of this folder.

7.2 Using the tool

The applications of this tool bode well in the healthcare ecosystem as the community slowly realizes the importance of AI-powered tools for the next generation of medical technology. Radiology has been one application that has been most welcoming of the new advances in technology as there is potential for substantial aid by reducing manual labor, increasing precision and freeing up the primary care physician’s time [1].

Yet, it is too early to take result the medical tool as gospel. For current cervical radiotherapy delineation tools, only 90% of the output is considered as acceptable for clinical use [13]. The remainder therefore has the potential to cause more harm than good if not checked properly. For example, overlap of a PTV onto an organ-at-risk may invoke a cascade of negative effects for the patient. A potential cause may be the lack of multivariate analysis, where an oncologist would need to consider a variety of data, whereas this model only considers a single point of evidence (results of an imaging modality). Clinicians can fall into the trap of automation-bias as AI becomes more common place in clinical environments [20]. However, many models of this age codify the existing bias in common cases,

which often will fail those patients who do not fit the expectations of the majority. Therefore, a degree of supervision required from physicians has to be established if this tool is to be used in practice. Oncologists will be required to reverse-engineer results of the ‘black-box’ to verify why a decision has been made. Secondly, the responsible party for incorrect decisions made by DL tools should also be determined [6].

Bibliography

- [1] Amisha et al. “Overview of artificial intelligence in medicine”. In: *Journal of Family Medicine and Primary Care* 8.7 (2019).
- [2] Richard Beare, Bradley Lowekamp, and Ziv Yaniv. “Image Segmentation, Registration and Characterization in R with SimpleITK”. In: *Journal of Statistical Software* 86.8 (2018), pp. 1–35. DOI: 10.18637/jss.v086.i08. URL: <https://www.jstatsoft.org/article/view/v086i08>.
- [3] Christopher M. Bishop and Hugh Bishop. *Deep Learning, Foundations and Concepts*. Springer, 2023.
- [4] Hester Breman. *NIfTI-1 Data Format*. URL: <https://nifti.nimh.nih.gov/nifti-1/documentation/nifti1diagrams/>.
- [5] Institute of Cancer Research and The Royal Marsden Hospital. *AMLART data*.
- [6] Zi-Hang Chen et al. “Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine”. In: *Cancer Communications* 41.11 (2021). URL: <https://onlinelibrary.wiley.com/doi/10.1002/cac2.12215>.
- [7] Bob Cox. *nifti-1 header field-by-field documentation*. URL: <https://nifti.nimh.nih.gov/pub/dist/src/niftilib/nifti1.h>.
- [8] *Data Protection Act 2018*. URL: <https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted>.
- [9] Michal Heker and Hayit Greenspan. *Joint Liver Lesion Segmentation and Classification via Transfer Learning*. Tech. rep. 2020. URL: <https://arxiv.org/pdf/2004.12352.pdf>.
- [10] Michele Larobina and Loredana Murino. “Medical Image File Formats”. In: *Journal of Digital Imaging* 27 (2013). URL: <https://link.springer.com/article/10.1007/s10278-013-9657-9>.
- [11] David B. Larson et al. “Ethics of Using and Sharing Clinical Imaging Data for Artificial Intelligence: A Proposed Framework”. In: (2020). URL: <https://pubs.rsna.org/doi/full/10.1148/radiol.2020192536>.
- [12] Xiangrui Li et al. “The first step for neuroimaging data analysis: DICOM to NIfTI conversion”. In: *Journal of Neuroscience Methods* 264 (2016). URL: <https://pubmed.ncbi.nlm.nih.gov/26945974/>.
- [13] Zhikai Liu et al. “Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy”. In: *Radiotherapy and Oncology* 153 (2020).
- [14] K. Mackay et al. *A Review of the Metrics Used to Assess Auto-Contouring Systems in Radiotherapy*. 2023. URL: [https://www.clinicaloncologyonline.net/action/showPdf?pii=S0936-6555\(23\)00021-3](https://www.clinicaloncologyonline.net/action/showPdf?pii=S0936-6555(23)00021-3).
- [15] Ying-Hwey Nai et al. “Comparison of metrics for the evaluation of medical segmentations using prostate MRI dataset”. In: *Computers in Biology and Medicine* 134 (2021), p. 104497. ISSN: 0010-4825. DOI: <https://doi.org/10.1016/j.compbiomed.2021.104497>. URL: <https://www.sciencedirect.com/science/article/pii/S0010482521002912>.

- [16] Stanislav Nikolov et al. "Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study". en. In: *J Med Internet Res* 23.7 (July 2021), e26151.
- [17] Sinno Jialin Pan and Qiang Yang. "A Survey on Transfer Learning". In: (2009). URL: <https://ieeexplore.ieee.org/abstract/document/5288526>.
- [18] Michael V Sherer et al. "Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review". en. In: *Radiother Oncol* 160 (May 2021), pp. 185–191. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9444281/>.
- [19] Nass SJ, Levit LA, and Gostin LO. "The Value and Importance of Health Information Privacy". In: (2009). URL: <https://www.ncbi.nlm.nih.gov/books/NBK9579/>.
- [20] Isabel Straw. "The automation of bias in medical Artificial Intelligence (AI): Decoding the past to create a better future". In: *Artificial Intelligence in Medicine* 110 (2020), p. 101965. ISSN: 0933-3657. DOI: <https://doi.org/10.1016/j.artmed.2020.101965>. URL: <https://www.sciencedirect.com/science/article/pii/S0933365720312306>.
- [21] Abdel Aziz Taha and Allan Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool". In: *BMC Medical Imaging* 15.1 (Aug. 2015), p. 29. ISSN: 1471-2342. DOI: 10.1186/s12880-015-0068-x. URL: <https://doi.org/10.1186/s12880-015-0068-x>.
- [22] Lisa Torrey and Jude Shavlik. *Transfer Learning*. Tech. rep. University of Wisconsin, 2009. URL: <https://ftp.cs.wisc.edu/machine-learning/shavlik-group/torrey.handbook09.pdf>.
- [23] The Royal Marsden NHS Foundation Trust. "Privacy Note". In: (2023). URL: https://rm-d8-live.s3.eu-west-1.amazonaws.com/d8live.royalmarsden.nhs.uk/s3fs-public/2023-10/T22020ac_Revised%20privacy%20policy_V1_AW_WEB.pdf.
- [24] Femke Vaassen et al. "Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy". en. In: *Phys Imaging Radiat Oncol* 13 (Dec. 2019), pp. 1–6.
- [25] wasserth et al. URL: <https://github.com/wasserth/TotalSegmentator>.
- [26] "What is Transfer Learning?" In: (). URL: <https://www.geeksforgeeks.org/ml-introduction-to-transfer-learning/>.
- [27] Varduhi Yeghiazaryan and Irina Voiculescu. "Family of boundary overlap metrics for the evaluation of medical image segmentation". In: *Journal of Medical Imaging* 5.1 (2018), pp. 015006–015006.