Imperial College London

MEng Individual Project

Department of Computing

Imperial College of Science, Technology and Medicine

Transfer Learning for Deep Learning Radiotherapy Planning

Author:
Anton Zhitomirsky

Supervisor:
Prof Ben Glocker

Second Marker:
Dr Thomas Heinis

May 31, 2024

Contents

Chapter 1

Introduction

1.1 Technical Context

1.2 Objectives and Contributions

1.3 Outline of Report

Chapter 2

Background

2.1 Clinical Context

This project will have its foundation for experimentation in a dataset provided by the Royal Marsden Hospital. We introduce this real-world clinical dataset, which has no exposure in academia and has uncommon and limited segmentation patterns, that will act as the pillar for justifying the success and transferability of ideas explored in this project to other medical domains. This dataset segments key anatomies and tumours which aid in radiotherapy planning for females with cervical cancer.

In this section, we discuss the clinical context behind cervical cancer in the population, the Hospital's pipeline for segmenting patients in preparation for radiotherapy treatment, and the Hospital's motivation for recruiting an AI tool to assist in its treatment pipeline.

2.1.1 Cervical Cancer

Cancer is a burden around the globe that has been a driver for almost one-sixth of the world's mortality in 2022 [1]. In females, cervical cancer makes up 25 countries' leading causes of cancer death, following breast cancer in 157 in 2022 [1]. Furthermore, the resulting maternal orphans from cancers affecting females experience health and education disadvantages throughout their lives [2]. Thankfully, cancer screening services provided by hospitals around Europe have been shown to decrease incidence and mortality rates of cervical cancer in women over the recent years [1], which inspires further complete clinical understanding of the disease. This motivation and the potential to pair with quality improvements offered by medical imaging models drive this research project to explore transfer knowledge in this field.

2.1.2 CT modality

The CT scan is a popular imaging modality in clinical environments because of its non-invasive ability to provide detailed images of the internal structures of the body. A series of X-ray devices are rotated around a specified body part, and computer-generated cross-sectional images are produced [3]. Whilst the scanner rotates, the table the patient lies on slowly moves up and down inside the tube to produce different cross-section images.

Hounsfield Units

(a) Muscle Window $(35, 55)$ [4] (b) Cancellous Bone Window $(300, 400)$ [5] (c) Fat Window $(-120, -90)$ [4]

Figure 2.1: Coronal view the same image slice of a CT image, with different window cropping (Patient id: 49, slice 251, axis 1)

The granularity or image slice thickness is decided by the operator or physician and ranges from 1mm to 10mm. Therefore, the precision along each axis creates a cube, or voxel which represents the value on a grid in three-dimensional space. The voxel values are measured in Hounsfield Units (HU) [6].

Contrary to natural images, where pixel values vary from 0 to 255 in 3 channels representing Red, Blue and Green, the Hounsfield scale is a quantitative scale for describing radiodensity where the image intensity reflects tissue type; each voxel intensity refers to specific tissue composition. The positive values are a consequence of more dense tissue with greater X-ray beam absorption, and negative values are less dense tissue with less X-ray beam absorption [7].

Therefore, becuase the HU scale is relative, different windows may be taken of a CT scan to highlight different tissues. Those voxels that lie within the window, are likely to be tissues of a specific classification. For example, disaplyed in Figure 2.1 we display 3 such windows, muscle, cancellous bone and fat.

2.1.3 Data Aquisition

Format

The data acquired through The Royal Marsden Hospital is presented as a set of 'Neuroimaging Informatics Technology Initiative' files (NIfTI) [3]. It serves as a lightweight alternative to other formats such as DICOM and eliminates ambiguity from spatial orientation information [8]. Libraries exist for handling these files, such as SimpleITK [9] which we use to read and manipulate the data in this project. The library reads, manipulates, and handles the image as a set of points in a grid occupying a physical region in space as defined by the metadata to remove ambiguity from the origin, size, spacing and so on that might vary between patient scans.

Training Data

Data was sampled for 100 varying female patients with similar types of cancer. With this data, come 7 relevant segmentation classes which contribute to radiotherapy planning. For the purpose of reproducibility, all delinated lablled data was lablled consistently by the Oncologists to improve chances that an AI model can learn cervical cancer patterns [10].

Finally, 10 hold-out data items were provided, which are patients with only the raw CT scan information with no labels.

Notes

Some notes contain clinical observations about each of the 100 labelled data pairs [10]. This small sample size of patients is also a good representation of the variability of data in the population. Because of the relatively small sample size, its important to be more acutely aware of the variability in the data. In particular, some observations are summarized.

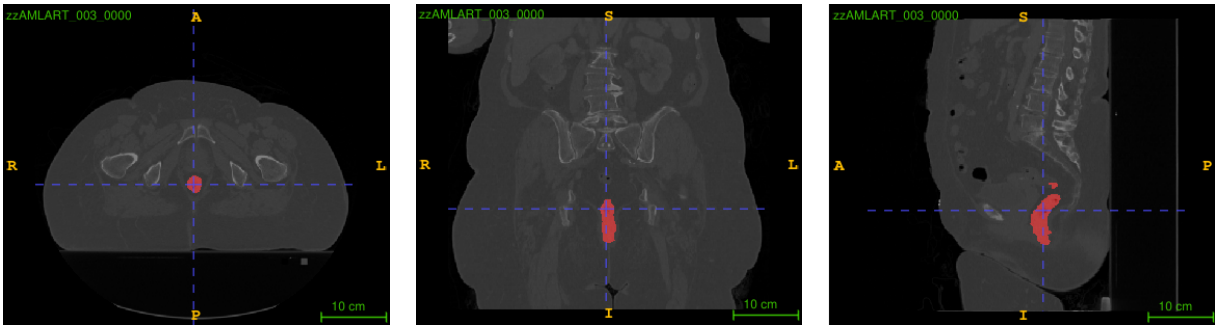
A common observation is that scans contain poor contrast. For patient 13, the note reads "no contrast - hard to see LNs" which is information crutial to determine for segmenting the Clinical Target Volume for Lymph Nodes (CTVn). Additionally, patients 9, 60, and 62 are examples that have "very large tumors", sometimes, these even shift into uncommon areas, with "sigmoid hanging into parametrium".

The notes are helpful to identify and diagnose some reasons for poor performance of the model which is characteristic of the high variability between patients.

2.1.4 Delineation classes

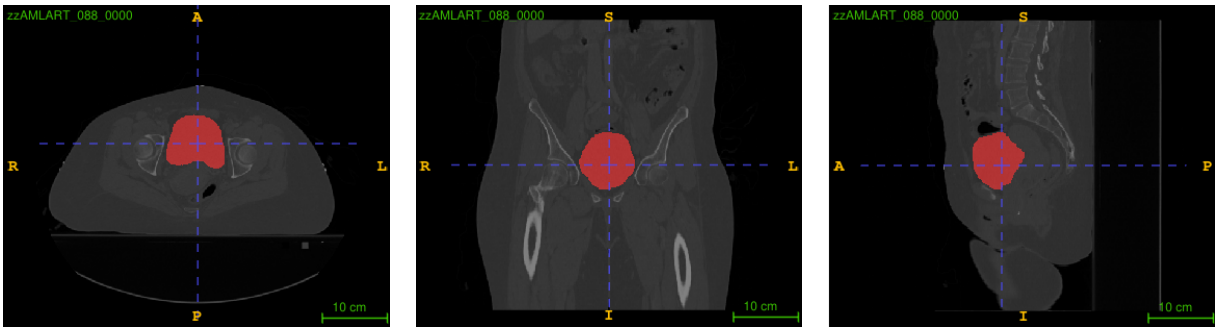
The clinicians at the Royal Marsden Hospital have provided segmentation labels for 5 high-priority classes of interest. These are the Bladder, Anorectum, CTVn, CTVp, and Parametrium.

Organs At Risk



(a) Axial (b) Coronal (c) Sagittal

Figure 2.2: Views of a segmented (in red) Anorectum of an arbitrary patient



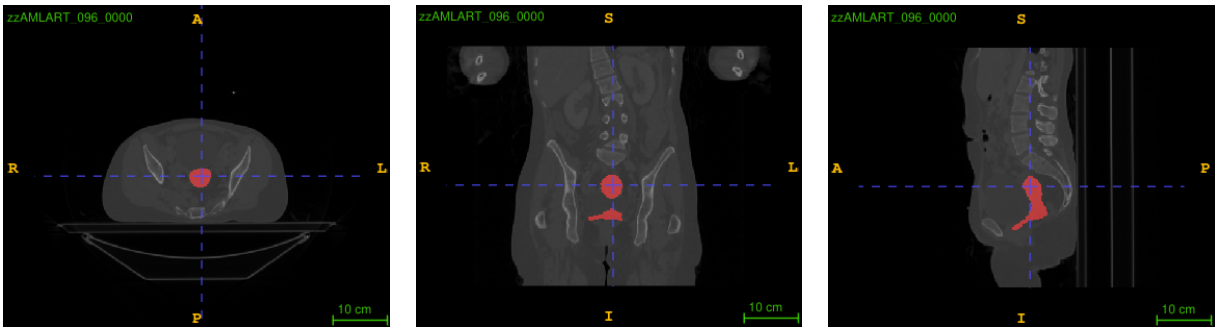
(a) Axial (b) Coronal (c) Sagittal

Figure 2.3: Views of a segmented (in red) Bladder 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 2.3(a)-2.3(c)) and the Anorectum (Figure 2.2(a)-2.2(c)). In particular, clinicians have identified that the Bladder may overlap with the CTVn (Section 2.1.4) and the Parametrium (Section 2.1.4).

CTVp



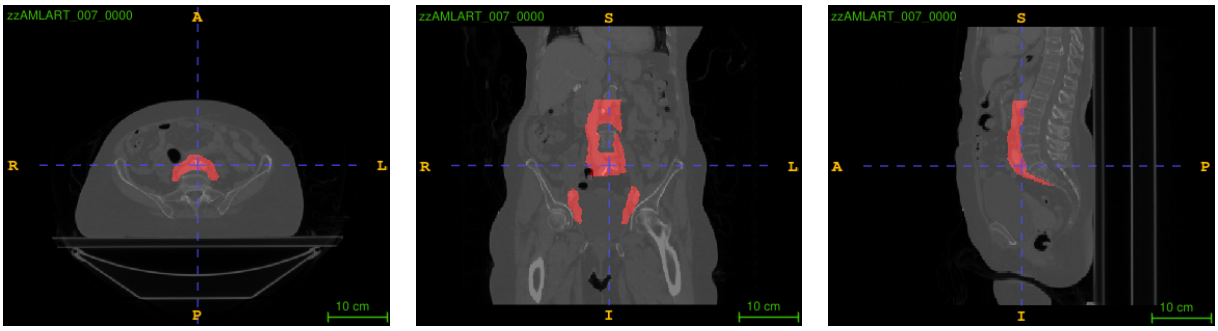
(a) Axial (b) Coronal (c) Sagittal

Figure 2.4: 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 2.4. This is the CTV where there may be local microscopic spread (uterus, cervix, upper vagina, primary tumour) [10]. 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 2.4.

CTVn



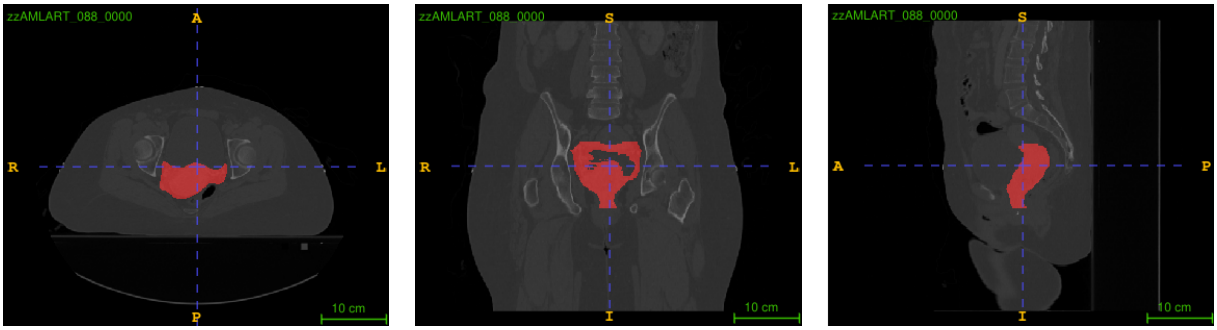
(a) Axial (b) Coronal (c) Sagittal

Figure 2.5: Views of a segmented (in red) CTVn of an arbitrary patient

The CTVn stands for Nodal Clinical Target Volume, see the example at Figure 2.5. 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 [10].

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 2.3.

Parametrium



(a) Axial (b) Coronal (c) Sagittal

Figure 2.6: 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 2.6. Drawn as a complete structure and editing back to the level of vagina to be included [10].

2.1.5 Rules

Notation of Structures

1. Let the Anorectum be denoted as A
2. Let the Bladder be denoted as B
3. Let the Cervix be denoted with C
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}
11. Let the Parametrium be denoted with P
12. Let the Uterus be denoted with U
13. Let the Vagina be denoted with V

Relationship between Structures

1. 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., $O^i = \{B^i, A^i, C_n^i, C_p^i, P^i\}$.
2. Let the overlap of two structures be denoted by the set intersect symbol $\cap$.
3. Let the joint area of two structures be denoted by the set union symbol $\cup$.

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):

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

$$\forall i, j \in \{C_n, C_p, A\} \text{ with } i \neq j, i \cap j = \emptyset \tag{2.1}$$

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

$$\forall i \in S, \quad P \cap S_i \neq \emptyset \quad \text{(Possibly)} \tag{2.2}$$

3. The Bladder may overlap with the CTVn.

$$B \cap C_n \neq \emptyset \vee B \cap C_n = \emptyset \tag{2.3}$$

4. The CTVp is defined as a compound structure containing:

$$C_p = \overbrace{C \cup G_p}^{\text{High Risk CTV}} \quad \cup \quad U \cup V \tag{2.4}$$

5. The CTVn is defined as a compound structure containing:

$$C_n = G_n \cup L_i \cup L_p + L_{pa} \tag{2.5}$$

2.2 Technical Context

2.2.1 AI in medical imaging

2.2.2 nnUNet

2.2.3 TotalSegmentator

2.2.4 UniverSeg

2.2.5 SAM

2.3 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.

Geometric measures are the most popular, a survey has found [11]. These measures compare an auto-contour to a ground-truth contour and return a relative score based on its performance.

2.3.1 Classification Based

Assesses if voxels within and outside the auto-contour have been correctly labelled [11]. 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 [12]) measures the proportion of predictions that were successfully correct. The Recall (also known as True Positive Rate [12]), 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".

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

2.3.2 Spatial Overlap Based

Similarly to Classification Based metrics in Section 2.3.1, an Overlap Based metric measures the extent of overlap between an auto-contour and a reference structure [11].

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

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

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 [11, 12, 13].

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

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 [14].

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

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

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.

2.3.3 Surface Based

Also commonly known as Boundary-Distance-Based Methods [15] compares the distance between two structure surfaces. These can be either maximum distance, average distance or distance at a set percentile of ordered distances [12].

A common example is the Haussdorf 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 [15]. Therefore, the better the HD metric, the smaller the value it returns. Here, the distance is taken by some norm, typically Euclidian distnace.

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

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 [15]. Therefore, we can calculate the average directed Haussdorf Distance.

2.3.4 Volume Based

Volume-based metrics consider only the volume of the segmentation [16, 11, 15]. However, due to its poor spatial descriptions it is more commonly used jointly with other metrics.

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

2.3.5 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.

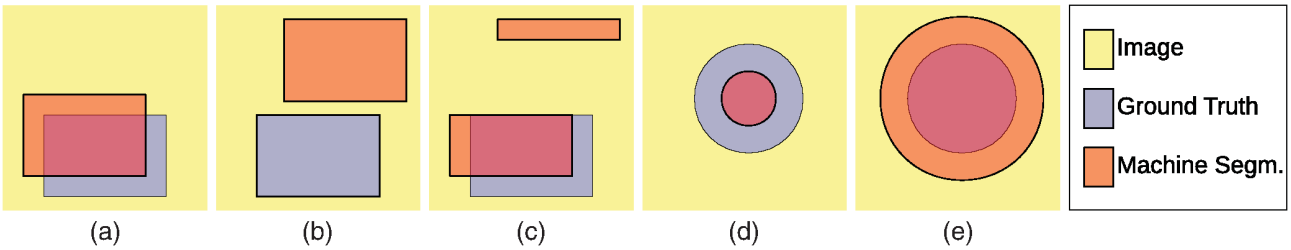


Figure 2.7: Figure from [15] illustrating cases of segmentation to aid with explanation of setbacks of certain evaluation metrics

- Classification Based (Section 2.3.1) and Spatial Overlap Based (Section 2.3.2) are similar; they are concerned with the number of correctly classified or misclassified voxels without taking into account their spatial distribution. Here, Figure 2.7(a) and Figure 2.7(c) would achieve similar results despite Figure 2.7(a) being locally bound to a better area.

- With Haussdorf Distance (Section 2.3.3) output segmentations generated by Figure 2.7(d) and Figure 2.7(e) will result in the same score, which is not favorable in a radiotherapy planning environment where an organ-at-risk is involved.

- Figure 2.7(b) would score flawlessly when using volumetric score estimation, however, it doesn't take into acount spatial placement, which makes this measurement rather poor when used individually.

2.3.6 Estimated Editing Based

☞ This is a quotation from this paper, [13], 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" [13].

Surface DSC

The study at [13] helped drive an initiative to combine aspects of Surface Based evaluation (Section 2.3.3) and Spatial Overlap Based evaluation (Section 2.3.2) into a Surface DICE. This assesses the specified tolerance instead of the overlap of the two volumes.

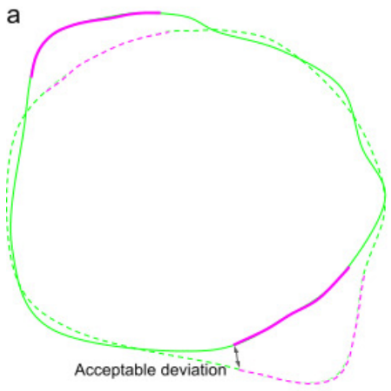


Figure 2.8: Taken from [17]. Illustrates the computation of the surface DICE, where the continuous line is the predicted surface and the dashed line is the ground truth. The black arrows show the maximum deviation tolerated without penalty; therefore, in pink is the unacceptable deviations and green otherwise.

We can formulate the Surface DSC score in a mathematical definition [13].

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

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.

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" [14]. 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 2.3.6)

☞ For future reference, *stack overflow discussion*

Implementation of surface DSC and APL: *source code*

2.3.7 Summary

This is why we settle at the Surface DSC (Section 2.3.6) 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 2.7(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 2.7(d) instead of Figure 2.7(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 [17].

Chapter 3

Methodology

3.1 Base-line nnUNet...

Chapter 4

Results

...

Chapter 5

Discussion

Here, discuss results from all 4 of the methods tested.

Chapter 6

Conclusion

Transfer works!

Chapter 7

Ethics

The lack of effort to protect the identities and confidentiality of patients during research projects may result in "stigma, embarrassment, and discrimination" [18] if the data is misused. This project involves very intimate and personal information of many female patients whose privacy must be established concretely before research is to take place.

7.1 Patient disclosures

Reserachers may collaborate with third-parties such as Imperial College London by providing anonymized data which may not be reverse engineered back to the patient. The collaborating hospital, 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. [19]. Formalities are also arranged with Imperial Collage's Medical Imaging team such as acting as "ethical data stewards" [20]. 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 [21]. 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 accepts the involvement of AI-powered medical tools. 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 [22].

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 [23]. 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 physician may base their final judgement subject to a multivariate analysis, which is contrary to the single image modality that this tool is based on. Therefore, the tool should be used as a second opinion rather than a primary source of information.

Clinicians can fall into the trap of automation-bias as AI becomes more common place in clinical environments [24]. 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 [25].

Bibliography

[1] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024. doi: https://doi.org/10.3322/caac.21834. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21834. pages 3

[2] Florence Guida, Rachel Kidman, Jacques Ferlay, Joachim Schüz, Isabelle Soerjomataram, Benda Kithaka, Ophira Ginsburg, Raymond B. Mailhot Vega, Moses Galukande, Groesbeck Parham, Salvatore Vaccarella, Karen Canfell, Andre M. Ilbawi, Benjamin O. Anderson, Freddie Bray, Isabel dos Santos-Silva, and Valerie McCormack. Global and regional estimates of orphans attributed to maternal cancer mortality in 2020. *Nature Medicine*, 28(12):2563–2572, Dec 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-02109-2. URL https://doi.org/10.1038/s41591-022-02109-2. pages 3

[3] Michele Larobina and Loredana Murino. Medical image file formats. *Journal of Digital Imaging*, 27, 2013. URL https://link.springer.com/article/10.1007/s10278-013-9657-9. pages 3, 4

[4] Lucas Haase, Jason Ina, Ethan Harlow, Raymond Chen, Robert Gillespie, and Jacob Calcei. The influence of component design and positioning on soft-tissue tensioning and complications in reverse total shoulder arthroplasty. *The Journal of Bone and Joint Surgery*, 12(4), 2024. doi: 10.2106/JBJS.RVW.23.00238. pages 4

[5] Herbert Lepor. *Prostatic Diseases*. W B Saunders Co Ltd, 1999. ISBN 978-0721674162. pages 4

[6] D.R. Dance, S. Christofides, A.D.A. Maidment, I.D. McLean, and K.H. Ng. *Diagnostic Radiology Physics*. International Atomic Energy Agency, 2014. pages 4

[7] DenOtter TD and Schubert J. *Hounsfield Unit*. StatPearls Publishing, Jan 2024. URL https://www.ncbi.nlm.nih.gov/books/NBK547721/. pages 4

[8] Xiangrui Li, Paul S. Morgan, John Ashburner, Jolinda Smith, and Christopher Rorden. The first step for neuroimaging data analysis: Dicom to nifti conversion. *Journal of Neuroscience Methods*, 264, 2016. URL https://pubmed.ncbi.nlm.nih.gov/26945974/. pages 4

[9] Richard Beare, Bradley Lowekamp, and Ziv Yaniv. Image segmentation, registration and characterization in r with simpleitk. *Journal of Statistical Software*, 86(8):1–35, 2018. doi: 10.18637/jss.v086.i08. URL https://www.jstatsoft.org/article/view/v086i08. pages 4

[10] Institute of Cancer Research and The Royal Marsden Hospital. Amlart data. pages 5, 6, 7

[11] K. Mackay, D. Bernstein, B. Glocker, and A. Taylor K. Kamnitsas. 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. pages 9, 10, 11

[12] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, Aug 2015. ISSN 1471-2342. doi: 10.1186/s12880-015-0068-x. URL https://doi.org/10.1186/s12880-015-0068-x. pages 9, 10

[13] Michael V Sherer, Diana Lin, Sharif Elguindi, Simon Duke, Li-Tee Tan, Jon Cacicedo, Max Dahele, and Erin F Gillespie. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother Oncol*, 160:185–191, May 2021. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9444281/. pages 10, 11, 12

[14] Femke Vaassen, Colien Hazelaar, Ana Vaniqui, Mark Gooding, Brent van der Heyden, Richard Canters, and Wouter van Elmpt. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol*, 13:1–6, December 2019. pages 10, 12

[15] Varduhi Yeghiazaryan and Irina Voiculescu. Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of Medical Imaging*, 5(1):015006–015006, 2018. pages 10, 11

[16] Ying-Hwey Nai, Bernice W. Teo, Nadya L. Tan, Sophie O'Doherty, Mary C. Stephenson, Yee Liang Thian, Edmund Chiong, and Anthonin Reilhac. Comparison of metrics for the evaluation of medical segmentations using prostate mri dataset. *Computers in Biology and Medicine*, 134:104497, 2021. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.104497. URL https://www.sciencedirect.com/science/article/pii/S0010482521002912. pages 11

[17] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, Christopher Kelly, Alan Karthikesalingam, Carlton Chu, Dawn Carnell, Cheng Boon, Derek D'Souza, Syed Ali Moinuddin, Bethany Garie, Yasmin McQuinlan, Sarah Ireland, Kiarna Hampton, Krystle Fuller, Hugh Montgomery, Geraint Rees, Mustafa Suleyman, Trevor Back, Cían Owen Hughes, Joseph R Ledsam, and Olaf Ronneberger. Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. *J Med Internet Res*, 23(7):e26151, July 2021. pages 12, 13

[18] Nass SJ, Levit LA, and Gostin LO. Beyond the hipaa privacy rule: Enhancing privacy, improving health through research. page 18, 2009. doi: 10.17226/12458. pages 18

[19] The Royal Marsden NHS Foundation Trust. Privacy note. URL https://rm-d8-live.s3.eu-west-1.amazonaws.com/d8live.royalmarsden.nhs.uk/s3fs-public/2023-10/T22020ac_Revisedprivacypolicy_V1_AW_WEB.pdf. pages 18

[20] David B Larson, David C Magnus, Matthew P Lungren, Nigam H Shah, and Curtis P Langlotz. Ethics of using and sharing clinical imaging data for artificial intelligence: A proposed framework. *Radiology*, 295(3):675–682, March 2020. doi: 10.1148/radiol.2020192536. pages 18

[21] *Data Protection Act 2018*. URL `https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted`. pages 18

[22] Amisha, Paras Malik, Monika Pathania, and Vyas Kumar Rathaur. Overview of artificial intelligence in medicine. *J Family Med Prim Care*, 8(7):2328–2331, July 2019. doi: 10.4103/jfmpc.jfmpc_440_19. pages 18

[23] Zhikai Liu, Xia Liu, Hui Guan, Hongan Zhen, Yuliang Sun, Qi Chen, Yu Chen, Shaobin Wang, and Jie Qiu. Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy. *Radiotherapy and Oncology*, 153:172–179, 2020. ISSN 0167-8140. doi: https://doi.org/10.1016/j.radonc.2020.09.060. pages 18

[24] Isabel Straw. The automation of bias in medical artificial intelligence (ai): Decoding the past to create a better future. *Artificial Intelligence in Medicine*, 110:101965, 2020. ISSN 0933-3657. doi: 10.1016/j.artmed.2020.101965. pages 19

[25] Zi-Hang Chen, Li Lin, Chen-Fei Wu, Chao-Feng Li, Rui-Hua Xu, and Ying Sun. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. *Cancer Communications*, 41(11), 2021. doi: 10.1002/cac2.12215. pages 19