Imperial College London

MEng Individual Project

Department of Computing

Imperial College of Science, Technology and Medicine

Transfer Learning for Deep Learning Radiotherapy Planning

Supervisor:
Prof Ben Glocker

Author:
Anton Zhitomirsky

Second Marker:
Dr Thomas Heinis

June 1, 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. The real-world clinical dataset segments key anatomies and tumours that aid in radiotherapy planning for females with cervical cancer. It has yet to gain exposure to the widespread segmentation challenges and has uncommon and limited segmentation patterns. This dataset will act as the pillar for justifying the success of the transferability of knowledge between medical domains.

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, an estimated 1 million maternal orphans who lose their mothers to cancer suffer long-term disadvantages in health and education [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 Radiotherapy Treatment

A mechanism available for cancer treatment involves radiation therapy. High beams of radiation energy are tuned to hone in to target cancerous cells in a clinically defined 'target area'. The cells killed by the energy experience interphase or proliferative death depending on the cell cycle stage. Death occurs when the damage to genetic material within the cell prevents it from dividing, or the cell's accumulation of genetic aberrations leads to a "mitotic catastrophe" [3]. In Europe, radiotherapy treatment was used on average for 70% of cases, with a curative rate of 40% [4, 5].

This death is characteristic of any cell subjected to high energy beams, placing much responsibility on the Oncologist to deliver an accurate treatment area so that healthy cells are unaffected; any adverse alteration of an organ's standard functionality may cause grave implications for the already compromised patient. Therefore, the precise and complex nature of the task is estimated to take oncologists 90-120 mins to delineate target areas for radiotherapy [6].

This time-consuming endeavour is never favourable for a patient already in a dangerous situation. For mid-low-income countries, where this may not be an available resource, this leaves them with a death rate 18 times that of a higher-income country [7].

2.1.3 CT modality

Development in high-resolution and high-contrast X-ray imaging (CT scans) machines has further incentivised radiotherapy treatment due to its noninvasive nature and ability to view patients' internal organs. X-ray devices rotate around a specified body part, and computer-generated cross-sectional images are produced [8]. Whilst the scanner rotates, the patient's table slowly moves up and down inside the tube to produce different cross-section images. The images show damaged and surrounding soft tissue, allowing physicians to propose clinical target volumes more accurately.

Hounsfield Units

(a) Muscle Window $(35, 55)$ [9] (b) Cancellous Bone Window $(300, 400)$ [10] (c) Fat Window $(-120, -90)$ [9]

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 operator or physician decides the granularity or image slice thickness, which ranges from 1mm to 10mm. Therefore, the precision along each axis creates a cube, or voxel, representing the value on a grid in three-dimensional space. The voxel values are measured in Hounsfield Units (HU) [11].

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 describing radiodensity. The image intensity reflects tissue type; each voxel intensity refers to a specific tissue composition. The positive values result from more dense tissue with greater X-ray beam absorption, and negative values are less dense tissue with less X-ray beam absorption [12].

Therefore, because the HU scale is relative, different windows may be taken for a CT scan to highlight different tissues. Those voxels within the window will likely be tissues of a specific classification. For example, as shown in Figure 2.1, we display three such windows: muscle, cancellous bone and fat.

2.1.4 Radiotherapy Planning

☞ Perhaps, include a graph to visualse target volumes visually

Oncologist use the CT scans to draw clinical volumes by combining their knowledge about the particular cancer to determine target structures, organs-at-risk structures, and areas where the cancer will likely spread to [13].

The first area is the macroscopic delineated area of the visible tumour area. This Gross Target Volume (GTV) has a high probability of containing the tumour. Secondly, the Clinical Target Volume (CTV) is derived to account for potential microscopic spread. The CTV will be an area at least as big as the GTV with an optional margin surrounding it containing a 'rind' of non-zero probability of tumour spread. Lastly, the Primary Target Volume (PTV) contains residual geometric uncertainties and safety margins surrounding the CTV, ensuring the radiotherapy dose gets delivered to the CTV [14, 15, 16, 17]. The PTV is a necessary extension of the CTV since geometric uncertainties are impossible and not advised to eliminate; after all, static scans are only estimations, subject to short-term organ misalignment, relative movement between structures of reference and tumours, partial volume effects and skewed anisotropic resolution [18].

In parallel, the Oncologist constantly considers critical healthy tissue structures that need to be preserved during irradiation. These are referred to as organs-at-risk (ORs). In some specific circumstances, adding a margin analogous to the PTV margin around an OR is necessary to ensure that the organ cannot receive a higher-than-safe dose; this gives a planning organ at risk volume [15].

2.1.5 International Guidelines

The final volumes have no internationally agreed-upon guidelines, which leaves it up to the interpretation of the oncologists and Hospitals to use their heuristics when drawing areas. This time-consuming process has high variability, causing it to suffer significantly from inter and intra-observer variability [16].

2.1.6 Data Aquisition

The Royal Marsden Hospital provides the dataset as a set of 'Neuroimaging Informatics Technology Initiative' files (NIfTI) [8]. It is a lightweight alternative to other formats such as DICOM and eliminates ambiguity from spatial orientation information [19]. Libraries exist for handling these files, such as SimpleITK [20], 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 defined by the metadata to remove ambiguity from the origin, size, spacing, and so on that might vary between patient scans.

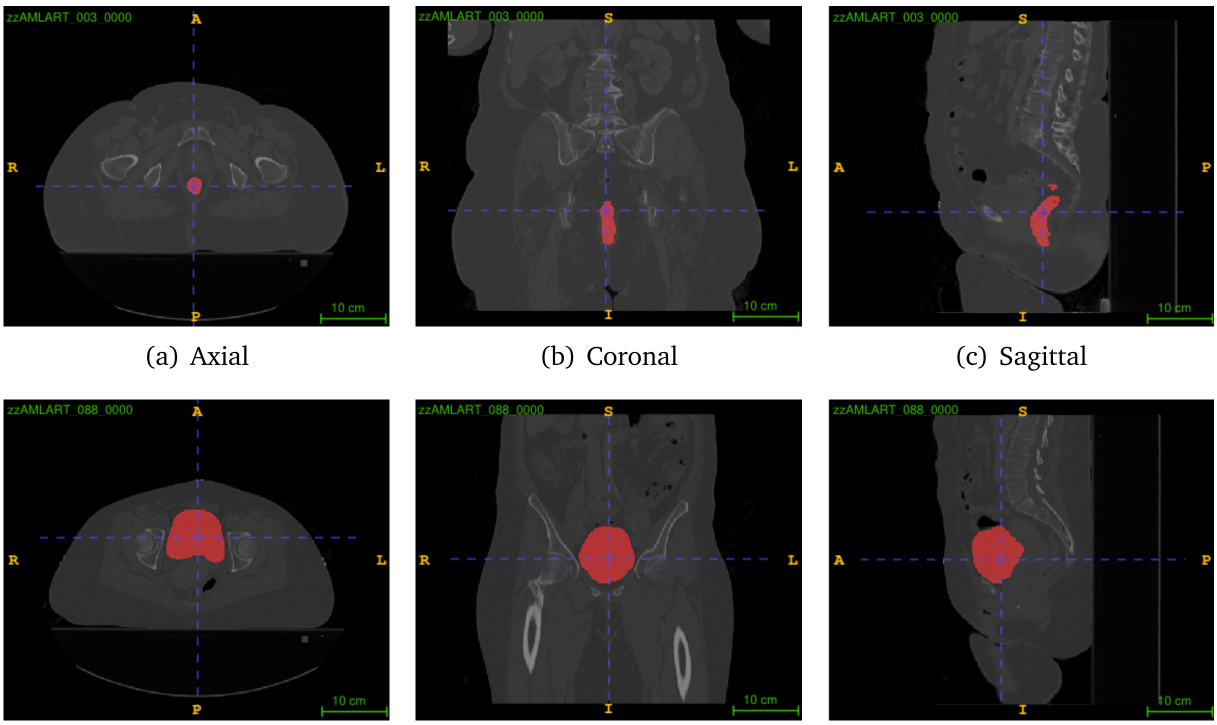
The training data provides one hundred female patients that have been diagnosed with similar types of cervical cancer. Each patient comes with seven relevant segmentation classes which contribute to radiotherapy planning. For reproducibility, all delineated anatomies were labelled consistently by the Oncologists to improve chances that an AI model can learn cervical cancer patterns [13].

Finally, the dataset comes with ten hold-out data items, which are patients with only the raw CT scan information without labels.

2.1.7 Delineation classes

The clinicians at the Royal Marsden Hospital have provided segmentation labels for seven high-priority classes of interest. These are the Bladder, Anorectum, CTVn, CTVp, Parametrium, Uterus, and Vagina. The function of these anatomies is irrelevant to this project and is left to the reader to research further. Note that for the sake of this report, the only deviation from clinical understanding is the CTV cancerous delineations.

Organs At Risk



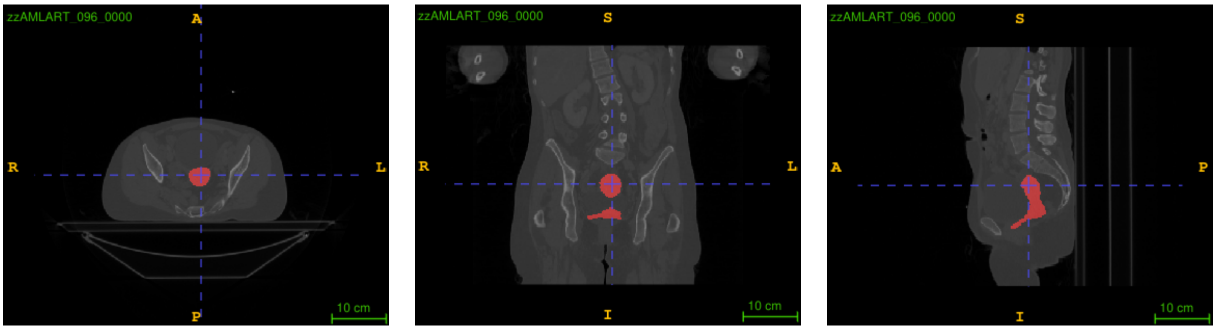
(a) Axial (b) Coronal (c) Sagittal

(d) Axial (e) Coronal (f) Sagittal

Figure 2.2: Views of the segmentation (in red) of the Anorectum (2.2(a)-2.2(c)) and the segmentation (in red) of the Bladder (2.2(d)-2.2(f)) of an arbitrary patient

An organ at risk is an organ that, despite being healthy, is substantially likely to be within the PTV. Any areas created around the area should actively avoid these organs because overlapping with them risks complicating treatment and compromising the health of functioning organs. The key supplied anatomies at risk are the Anorectum (Figure 2.2(a)-2.2(c)) and the Bladder (Figure 2.2(d)-2.2(f)).

CTVp



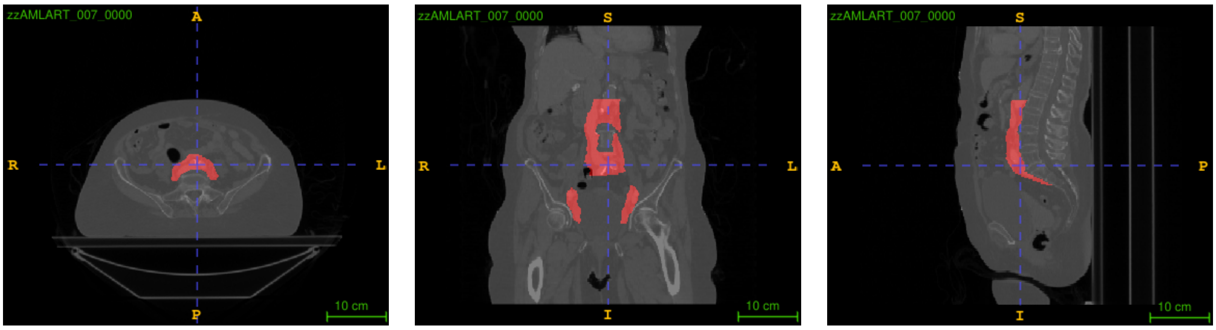
(a) Axial (b) Coronal (c) Sagittal

Figure 2.3: 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 2.3. This is an extension of the CTV definition (Section 2.1.4) where there may be local microscopic spread (uterus, cervix, upper vagina, primary tumour) [13]. This is the area that contains the tumour.

The CTVp is not an organ in a body but rather an area comprised of other components formed by joining other structures. The CTVp is an area defined in Equation **??**.

CTVn



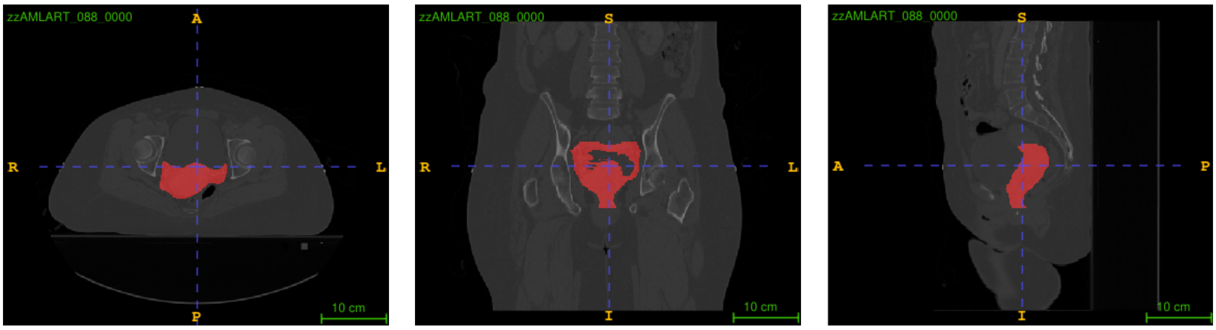
(a) Axial (b) Coronal (c) Sagittal

Figure 2.4: Views of a segmented (in red) CTVn of an arbitrary patient

The CTVn stands for Nodal Clinical Target Volume; see the example at Figure 2.4. Once again, this is an extension of the CTV definition (Section 2.1.4), with the enhanced condition that 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 [13].

Similarly to CTVp, this is a compound area with three groups of lymph nodes. In clinical practice, the number of these groups in the CTV varies in each patient, depending on the advanced disease. The CTVn is an area defined in Equation 2.3.

Parametrium



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

Figure 2.5: 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 2.5. The Parametrium is drawn as a complete structure and edited back to the level of the vagina to be included [13].

Vagina and Uterus

☞ Insert similar figure here

The final structures are the Vagina and Uterus, their clinical significance is to help define encapsulating structures like the CTVn (see Section 2.1.8).

2.1.8 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. If we want to talk about a specific patient, we should use the super-script notation to differentiate patients, e.g., B^i for the Bladder of patient 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 seven 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.

$$\forall i \in S, \quad P \cap S_i \neq \emptyset \vee P \cap S_i = \emptyset \tag{2.2}$$

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:

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

However, since we are never explicitly provided with the segmentation maps for the Cervix C and the GTVp G_p, we cannot use as strong of a definition as above. Instead, we operate on the assumption that the union of the Uterus and Vagina is at least as big as the CTVp.

$$U \cup V \subseteq C_p \tag{2.5}$$

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

$$C_n = G_n \cup L_i \cup L_p + L_{pa} \tag{2.6}$$

Similarly, we are not provided segmentations for these areas, therefore, operating under no clinical knoweldge apart from the provided, cannot make any claims as to the composition of the CTVn.

2.1.9 Motivation in AI

The medical sector has been a hotbed for AI research since researchers realised they could apply Convolutional Neural Networks (Section 2.2.1) to medical image data. A branch of research dedicated itself to segmentation, which involves labelling individual pixels in the image according to which object or class they belong to. In dense classification, a model assigns every pixel to a specific class. Relevant to the direction of this project is determining the precise location and extent of organs or certain types of tissue, like ORs, CTV volumes, or other anatomies.

The key objective of models trained for delineating target structures for this project is to see if an AI model can learn cervical cancer CTV pattern detection. The decision is complex as clinicians use information beyond the CT-imaging modality, such as how far along the tumour has progressed, and other clinical intuition to make proper judgements about the CTV volumes. Therefore, with this information missing from AI models, it is likely to misjudge target volumes, and a clinician will have to select which components of the CTV are required. However, a clinician will likely benefit from the time saved and improved consistency with the planning process if a trained model can produce the substructures required within the CTV that a clinician can review [13].

2.2 Technical Context

Before popularising machine learning algorithms, strict and convoluted rule sets defined algorithms. These heuristically defined algorithms needed to be more scalable to complex problem sets and were often complicated or confusing to maintain. The typical task, however, does not warp easily into human intuition.

Algorithms began to emerge that fell into the classification of neural network models. Observations were rephrased and morphed into vectorised inputs, where each constituent of the vector represented a particular feature of the observation. For instance, the California Housing dataset [21] contains an input of 9 features (longitude, latitude, number of bathrooms, ...) and one target feature (house price). After pre-processing and strategising, this input would be vectorised and fed into a network with several layers. Within each layer, sets of tunable parameters would optionally change the number of features and learn a relationship between parameters using weights until the 1×9 vector finally translates into a scalar value indicating the house price. After many repetitions of learning from examples, the model would learn an approximation to the solution. These models were termed Multi-Layered Perceptrons (MLPs).

At the same time, J. Hull, sponsored by the United States Postal Service, published a Database for Handwritten Text Recognition with the incentive of providing an extensive dataset of 300-pixel images of characters of variable writing mediums, isolation, overlap, and neatness to aid research efforts in developing accurate digit classification algorithms [22].

The current machine learning approach didn't work well with image data. Up until that point, rich structures such as images were neglected, and matrices of gray-scale images were mutilated into flat vectors. This approach was necessary to feed the flattened image representation through the MLP. The issue with vectorising an image is that it loses its spatial context-driven awareness. Yann LeCunn et al. [23] was the first to propose a convolution-based approach, which preserved the input in its 2D glory by applying a set of slideable convolutional blocks and morphing it into intermediate feature maps. This first convolutional network spawned a vicious flurry of convolutional architectures, which followed in their footsteps.

2.2.1 Convolutions for Segmentation

The Fully Connected Neural Network (FCN) was the first attempt to adapt popular architectures for image classification. Previously, convolutions would reduce input image feature vectors into non-spatial classification outputs. However, this paper 'convolutionalized the pipeline to provide a heatmap of segmented objects within the image [24]. The heatmap would describe in a 2D feature the location of each class.

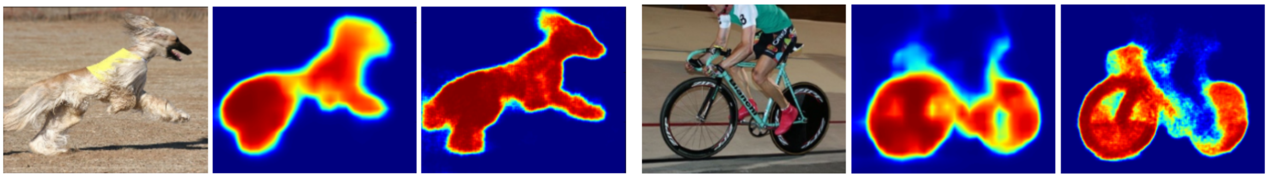


Figure 2.6: Comparison of upsampling a base image using FCN [24] and the VGG-16-based DeconvNet [25, 26] architectures.

Trivial upsampling through de-convolutional layers allows the heatmap to translate back to the original size. The trivial process would produce a largely inaccurate segmentation with much room for improvement. Therefore, similarly to learning the downsampling, the model learnt to upsample low-level heatmap representations [25]. This way, the deconvolutional network also became "a key component for precise object segmentation," which improved the base upsampling provided by the FCN. This conclusion is shown in Figure 2.6.

2.2.2 UNet

At first a

2.2.3 nnUNet

2.2.4 TotalSegmentator

2.2.5 UniverSeg

2.2.6 SAM

2.3 Evaluation Metrics

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

If we were attempting to fit a model onto a line in 2D space, the performance of our model would be the total minimum distance between each point and the prediction. Our objective would be to drive the model's distance metric as close to 0 without overfitting. Here, the points act as a 'ground truth', alternatively referred to as the gold standard, which represents the actual measured value.

The reasoning above extends to 3D and 2D in a segmentation context with variants to measure other quantities, like the minimum distance between prediction and truth or the extent of volume overlap between the two. These are examples of geometric measures, which Mackay et al. has found to be the most popular measure in segmentation tasks [27].

2.3.1 Classification Based

Assesses if voxels within and outside the auto-contour have been correctly labelled [27]. To begin, we define 'positive' to mean that the voxel selected indeed needs radiotherapy treatment and 'negative' to mean that the voxel classifies as healthy.

A standard measure of classification is accuracy. It measures the total number of correct predictions vs. the total predictions it made. However, more than this measure is needed to fully capture a model's bias because it does not tell the whole story with class-imbalanced data when there is no 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 [28]) measures the proportion of successfully correct predictions. The Recall (also known as True Positive Rate [28]), on the other hand, "measures the portion of positive voxels in the ground truth that is 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 [27].

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

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

A specific case of this equation with $\beta = 1$ is mathematically equivalent to the DICE Similarity Coefficient. A review found that DICE is the most popular evaluation metric amongst 2021 studies [27, 28, 29].

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

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 in classical machine learning; it can 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 argument extends to a Volumetric DSC by considering the above in all three dimensions [30].

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 than the DICE (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 [31] compares the distance between two structure surfaces. These can be maximum, average or distance at a set percentile of ordered distances [28].

A typical example is the Haussdorf Distance. Here, a directed distance metric is the maximum distance from a point in the first set to the nearest point in the other between two individual voxels [31]. Therefore, the better the HD metric, the smaller the value it returns. Here, the distance is typically Euclidian distance.

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

The HD is generally sensitive to outliers; therefore, direct HD application gives uninspiring results because noise and outliers are common in medical segmentations [31]. Therefore, we can calculate the average directed Haussdorf Distance.

2.3.4 Volume Based

Volume-based metrics consider only the volume of the segmentation [32, 27, 31]. However, its poor spatial descriptions make it 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 others. To decide which segmentation is best, we can list some challenging scenarios.

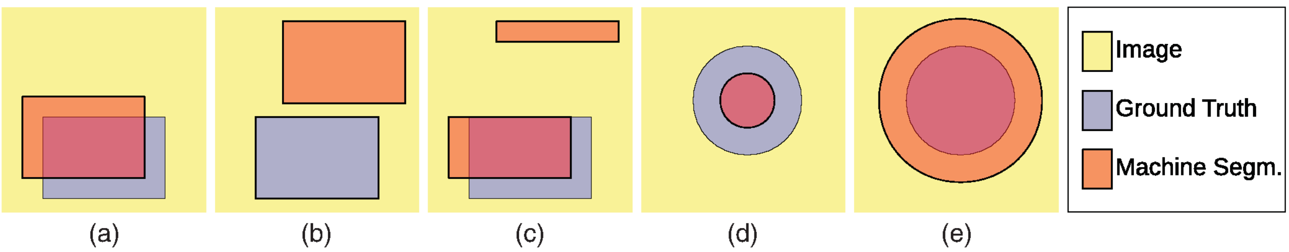


Figure 2.7: Figure from [31] illustrating cases of segmentation to aid with explanation of setbacks of certain evaluation metrics

- 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 favourable 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 does not consider spatial placement, making this measurement poor when used individually.

2.3.6 Estimated Editing Based

☞ This is a quotation from this paper, [29], however, it is referencing a paper of its own. Shall I reference the original paper, or are 'linked' references OK?

Selecting a measurement that can reflect a clinician's acceptability score is difficult. A study found a lack of correlation between a geometric index and expert evaluation, with the JAC score having a 13% False Positive Rate. The study's conclusion summarised that scores such as JSC and volumetric DSC "provide limited clinical context and correlation with clinical or dosimetric quality" [29].

Surface DSC

The study at [29] 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 which assesses the specified tolerance instead of the overlap of the two volumes.

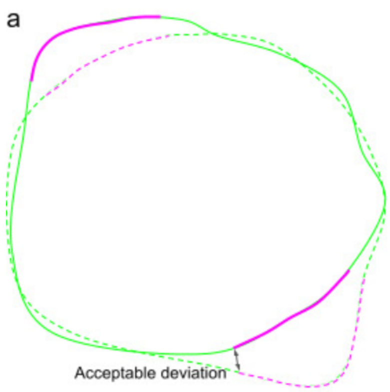


Figure 2.8: Taken from [33]. Illustrates the computation of the surface DICE, where the continuous line is the predicted surface, and the dashed line is the ground truth. The black arrows show the maximum deviation tolerated without penalty; therefore, in pink are the unacceptable deviations and green otherwise.

We can formulate the Surface DSC score in a mathematical definition [29].

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

This definition measures 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

Similarly, the APL score predicts "the path length of a contour that has to be added" [30]. APL 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 the 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 this project, we shall select an evaluation measurement more biased towards conservative boundary estimates not to touch the organs at risk. The clinician's review pipeline, in part, influenced this choice; it would be 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 [33].

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" [34] if the data is misused. This project involves the intimate and personal information of many female patients whose privacy must be protected before research occurs.

7.1 Patient disclosures

Researchers may collaborate with third parties, such as Imperial College London, by providing anonymised data that third parties cannot reverse-engineer to identify the patient. The collaborating hospital, The Royal Marsden Hospital, does not 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/anonymised" data may be used. [35]. Imperial College's Medical Imaging team also arranges formalities, such as acting as "ethical data stewards" [36].

The MIRA team acts as responsible data stewards by storing anonymised data within a folder on the college network. They received all provided data in the `NIfTI` file format, which discloses no personally identifiable information, as defined by the GOV website [37]. Specific access rights limit data availability in this folder, ensuring security measures. Moreover, taking the data outside this folder reduces individual patient risk through the exchange of de-identified data.

Without such disclosure, anonymisation, and a security guarnatee of the data, patients may be reluctant to provide candid and complete disclosures of their sensitive information, even to physicians, which may prevent a complete diagnosis if their data is not maintained anonymously.

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 is one application that has been most welcoming of the new technological advances as there is potential for substantial aid by reducing manual labour, increasing precision and freeing up the primary care physician's time [38].

However, it is too early to take the results of the medical tool as gospel. For current cervical radiotherapy delineation tools, only 90% of the output is acceptable for clinical use [39].

Therefore, The remainder can potentially cause more harm than good if not checked properly. For example, the overlap of a PTV with an organ-at-risk may invoke a cascade of adverse effects for the patient. The remaining 10% of outputs may score incorrectly because the model uses a single modality, but physicians may base their final judgement on a multivariate analysis. Therefore, clinicians should use the tool as a second opinion rather than a primary source of information. Otherwise, an ethical dilemma of establishing the responsible party for incorrect decisions made by DL tools should also be determined [40].

Clinicians can fall into the trap of automation bias as AI becomes more commonplace in clinical environments [41]. However, many models of this age codify the existing bias in common cases, which often will fail those patients who do not fit the majority's expectations.

Therefore, before integrating tools into workflows, a committee must establish the degree of supervision required from physicians if this tool is to be used in practice. Currently, oncologists will be required to reverse-engineer the 'black box' results to verify why a decision has been made.

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] Yunfei Jiao, Fangyu Cao, and Hu Liu. Radiation-induced cell death and its mechanisms. *Halth Phys.*, 2022. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9512240/pdf/hpj-123-376.pdf`. pages 3

[4] Rajamanickam Baskar, Kuo Ann Lee, Richard Yeo, and Kheng-Wei Yeoh1. Cancer and radiation therapy: Current advances and future directions. 2012. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3298009/`. pages 3

[5] Mareike K. Thompson, Philip Poortmans, Anthony J. Chalmers, Corinne Faivre-Finn, Emma Hall, Robert A. Huddart, Yolande Lievens, David Sebag-Montefiore, and Charlotte E. Coles. Practice-changing radiation therapy trials for the treatment of cancer: where are we 150 years after the birth of marie curie? *British Journal of Cancer*, 119 (4):389–407, Aug 2018. ISSN 1532-1827. doi: 10.1038/s41416-018-0201-z. URL `https://doi.org/10.1038/s41416-018-0201-z`. pages 3

[6] Zhikai Liu, Xia Liu, Bin Xiao, Shaobin Wang, Zheng Miao, Yuliang Sun, and Fuquan Zhang. Segmentation of organs-at-risk in cervical cancer ct images with a convolutional neural network. *Physica Medica*, 69:184–191, 2020. ISSN 1120-1797. doi: https://doi.org/10.1016/j.ejmp.2019.12.008. URL `https://www.sciencedirect.com/science/article/pii/S1120179719305290`. pages 4

[7] William Small, Monica A. Bacon, Amishi Bajaj, Linus T. Chuang, Brandon J. Fisher, Matthew M. Harkenrider, Anuja Jhingran, Henry C. Kitchener, Linda R. Mileshkin, Akila N. Viswanathan, and David K. Gaffney. Cervical cancer: A global health crisis. *Cancer, An international interdisciplinary journal of American Cancer Society*, 123(13), 2017. URL `https://acsjournals.onlinelibrary.wiley.com/doi/10.1002/cncr.30667`. pages 4

[8] 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 4, 5

[9] 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

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

[11] 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

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

[13] Institute of Cancer Research and The Royal Marsden Hospital. Amlart data. pages 5, 6, 7, 8, 10

[14] C. F. Njeh. Tumor delineation: The weakest link in the search for accuracy in radiotherapy, 2008. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2772050/. pages 5

[15] Neil G Burnet, Simon J Thomas, Kate E Burton, and Sarah J Jefferies. Defining the tumour and target volumes for radiotherapy, 2004. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1434601/pdf/ci040153.pdf. pages 5

[16] Hui Lin, Haonan Xiao, Lei Dong, Kevin Boon-Keng Teo, Wei Zou, Jing Cai, and Taoran Li. Deep learning for automatic target volume segmentation in radiation therapy: a review. *Quant Imaging Med Surg*, 11(12):4847–4858, December 2021. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8611469/. pages 5

[17] David Bernstein, Alexandra Taylor, Simeon Nill, and Uwe Oelfke. New target volume delineation and ptv strategies to further personalise radiotherapy, 2021. pages 5

[18] Marcel van Herk. Errors and margins in radiotherapy. *Seminars in Radiation Oncology*, 14(1):52–64, 2004. ISSN 1053-4296. doi: https://doi.org/10.1053/j.semradonc.2003.10.003. URL https://www.sciencedirect.com/science/article/pii/S1053429603000845. High-Precision Radiation Therapy of Moving Targets. pages 5

[19] 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 5

[20] 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 5

[21] R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. doi: 10.1016/S0167-7152(96)00140-X. pages 10

[22] J.J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440. pages 10

[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. pages 10

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. Technical report, Berkley, 2015. URL `https://arxiv.org/pdf/1411.4038.pdf`. pages 10, 11

[25] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation, 2015. pages 11

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. pages 11

[27] 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 11, 12, 13

[28] 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 12, 13

[29] 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 12, 14

[30] 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 12, 15

[31] 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 13

[32] 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 13

[33] 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 14, 15

[34] 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 20

[35] 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 20

[36] 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 20

[37] *Data Protection Act 2018*. URL `https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted`. pages 20

[38] 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 20

[39] 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 20

[40] 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 21

[41] 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 21