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

June 7, 2024

Abstract

Contents

Chapter 1

Introduction

1.1 Technical Context

1.2 Objectives and Contributions

1.3 Outline of Report

The report will first discuss the relevant background knowledge required for this project in Chapter 2. Within, we provide a high-level overview of anatomies and the clinical context (Section 2.1) as well as core existing academic knowledge in the Computer-assisted vision in the Medical Imaging field (Section 2.2-3.5).

Then, Chapter 3 describes the experiments used to evaluate how well different architectures transfer knowledge into the target domain, as well as an overview of the results (Chapter 4) as well as a discussion of their implications (Chapter 4).

Chapter 2

Motivation

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 for 157 countries 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. Paired with quality improvements offered by medical imaging models, the motivation for total control over cervical cancer 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

High-resolution and high-contrast CT machines have further benefited cancer treatment due to their 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) 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). White areas represent high-density tissues, and black areas represent low-density tissues within the window range.

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 (white) result from more dense tissue with greater X-ray beam absorption, and negative values (black) 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 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 cervical cancer. For reproducibility, all delin-

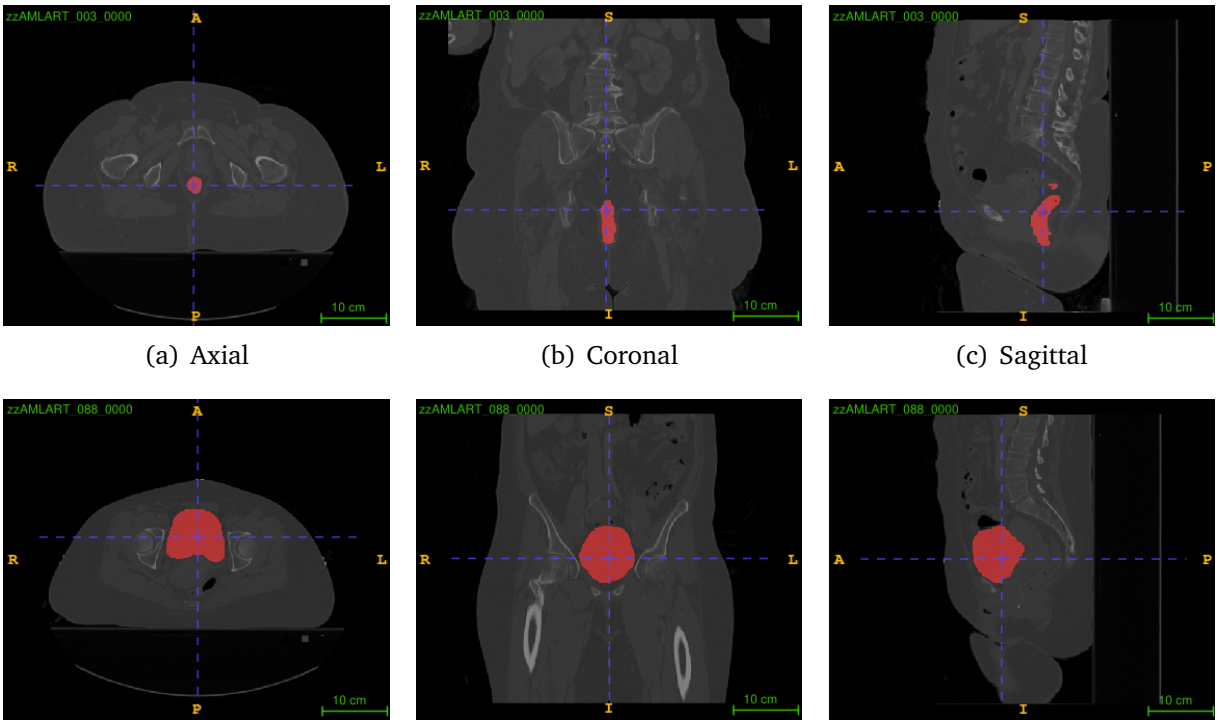
eated anatomies were labelled consistently by the Oncologists to improve chances that an AI model can learn cervical cancer patterns [13].

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 regions of interest (roi). 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.

Organs At Risk



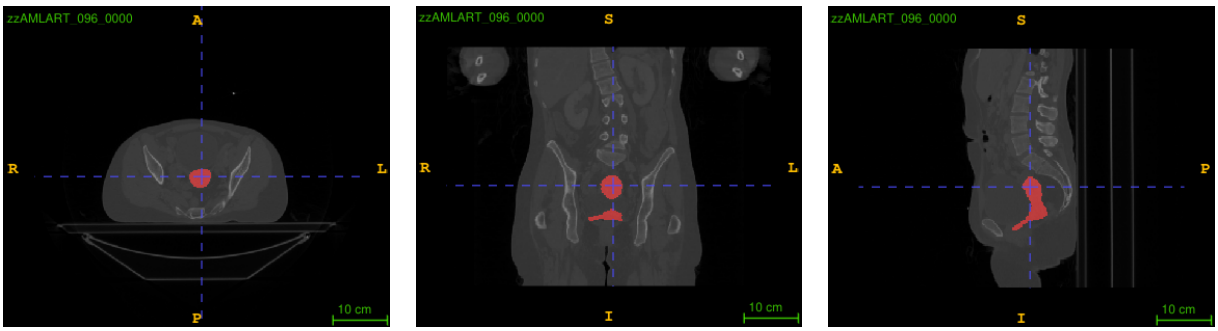
(a) Axial (b) Coronal (c) Sagittal

(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

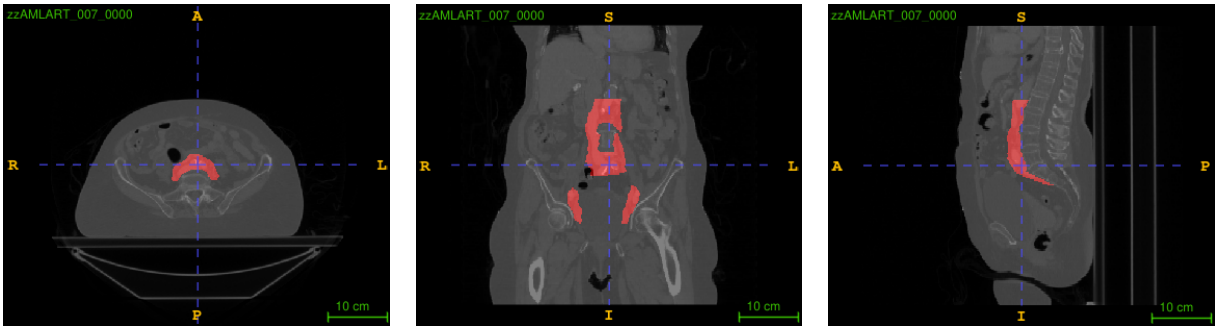


(a) Axial (b) Coronal (c) Sagittal

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 area comprised from areas where there may be local microscopic spread (uterus, cervix, upper vagina, primary tumour) [13].

CTVn



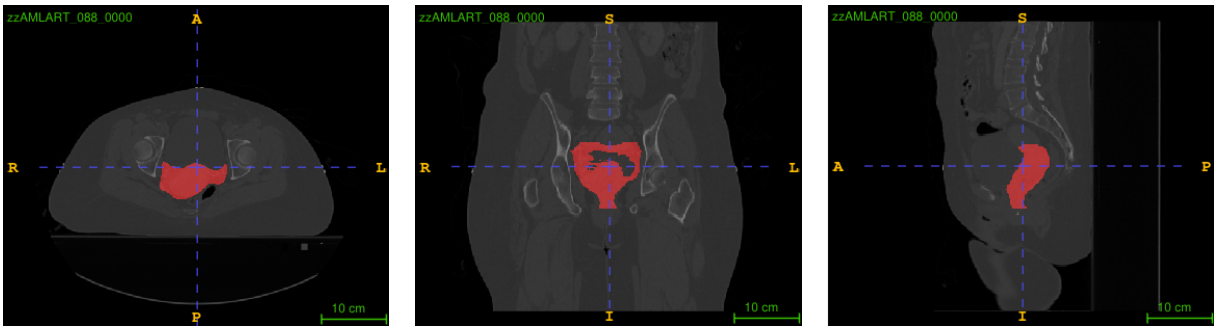
(a) Axial (b) Coronal (c) Sagittal

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. This CTV surrounds areas that may contain 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. However, in contrast to the CTVp, this area is drawn depending on the development of the disease.

Parametrium

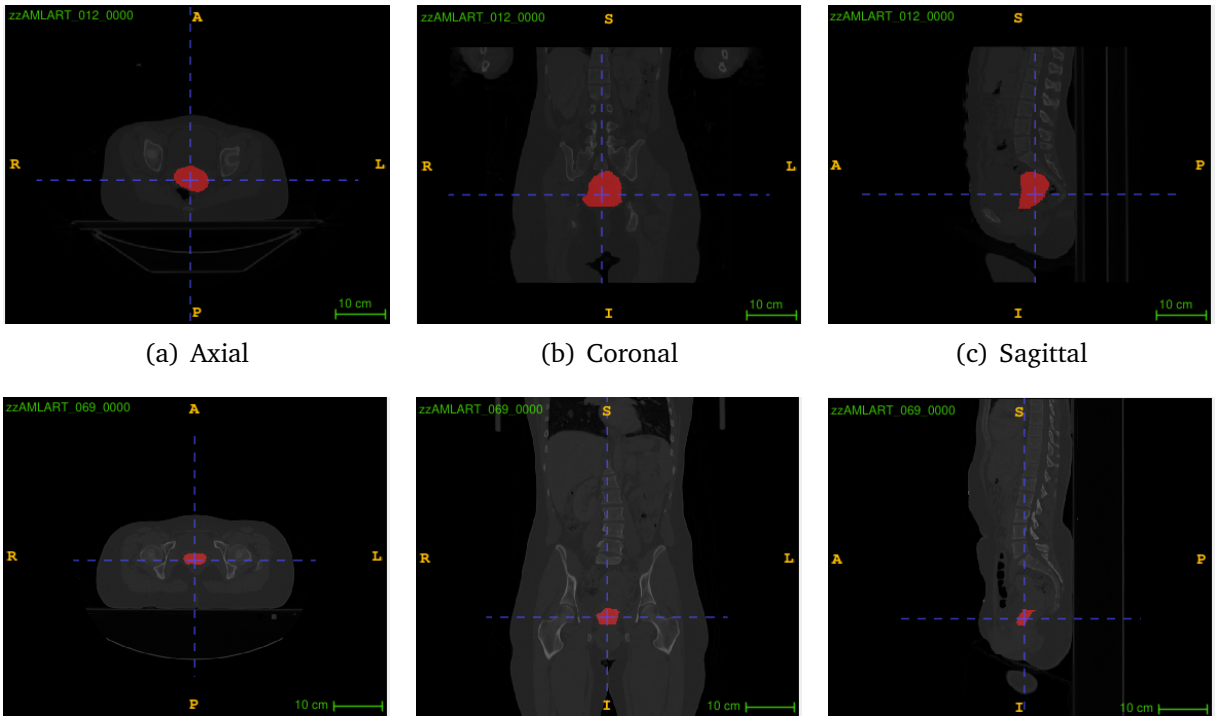


(a) Axial (b) Coronal (c) Sagittal

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



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

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

Figure 2.6: Views of the segmentation (in red) of the Uterus (2.6(a)-2.6(c)) and the segmentation (in red) of the Vagina (2.6(d)-2.6(f)) of an arbitrary patient

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

Let us represent each organ anatomy as the first letter of its name, specifically: (A)norectum, (B)ladder, (C)ervix, (P)arametrium, (U)terus, (V)agina. Further, define:

1. The CTVn and CTVp as C_n and C_p respectively
2. The GTVn and GTVp as G_n and G_p respectively
3. The Pelvic, Common and Para-aortic Lymph Node as L_p, L_c, and L_{pa} respectively

Relationship between Structures

1. Let the overlap of two structures be denoted by the set intersect symbol $\cap$.
2. 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.

$$\forall i \in S, \quad (P \cap S \neq \emptyset) \vee (P \cap S = \emptyset), \quad \text{where } S = \{A, B, C, C_n, C_p, U, V\} \tag{2.2}$$

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 Machine learning for image segmentation

Before popularising machine learning algorithms, strict and convoluted rule sets defined algorithms. These heuristically defined algorithms struggled to scale to complex problems 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).

2.2.1 Image Segmentation

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.

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 images of characters of variable writing mediums, isolation, overlap, and neatness to aid

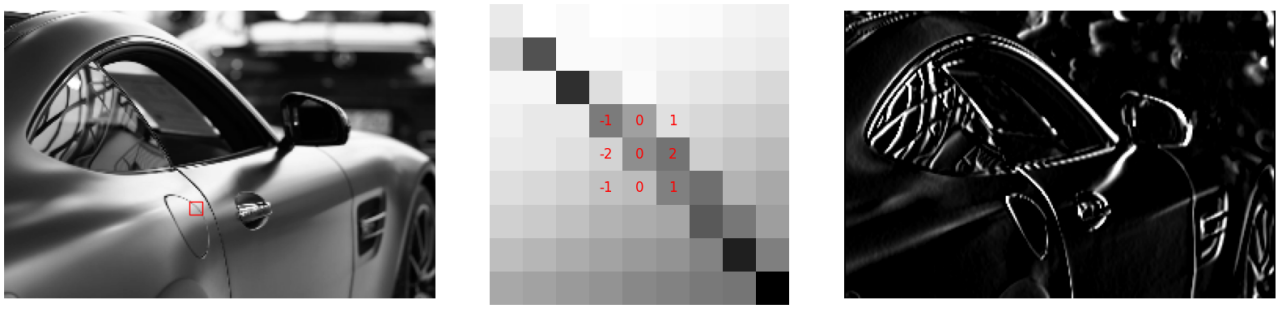


Figure 2.7: Example application of a convolution. From left to right, the input image with a region outlined in a red box, the boxed region magnified with a convolutional (sobel) filter being applied to a part of the magnified region, and lastly the output after the filter has passed over the entire image. The output represents a new feature map encoding features of the original (input) feature.

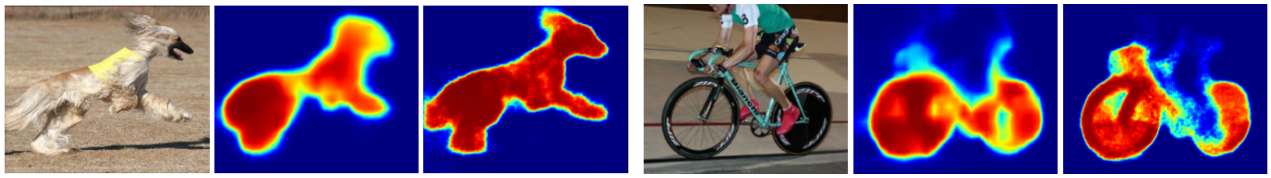


Figure 2.8: Comparison of upsampling a base image using FCN [24] and the VGG-16-based DeconvNet [25, 26] architectures.

research efforts in developing accurate digit classification algorithms [22]. Yann LeCunn et al. [23] used this database to propose the first Convolutional Neural Network (CNN) for image processing, which preserved the input in its 2D glory by applying convolutions.

Convolutional layers are rectangular blocks that are recipes for translating an image (the input feature). The algorithm centres the block over a specific pixel and uses a square radius of neighbouring pixels. It multiplies and sums the pixels along the corresponding pixel positions according to the recipe to produce a transformed resulting pixel that encodes the reference pixel's information and the surrounding receptive field around it. Figure 2.7 demonstrates this concept. Specifically, the middle tile which shows an example 3×3 convolutional filter being applied to a zoomed in part of the image. This filter slides across the entire image and encodes the entire image which produces the output on the right of Figure 2.7.

The CNN operates by having multiple learnable convolutional filters stacked on top of each other. The values within the square convolutional filter are *learned* during training; the values learnt are those that, in combination with the layers before and after, encode the image's features according to the training objective the best. When stacked with many other convolutional filters that piggyback off the encoded features produced by filters before it, this allows for dense feature representations that encode the entire image.

This first convolutional network spawned a vicious flurry of convolutional architectures, which followed in their footsteps. The Fully Connected Neural Network (FCN) adapted the architecture used by LeCunn et al. [23] for segmentation applications. 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.

Trivial upsampling through de-convolutional layers allows the heatmap to translate back to

the original size. This 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 2.8.

The strategy of downsampling and upsampling for image segmentation is a common theme amongst many segmentation architectures.

2.2.2 UNet

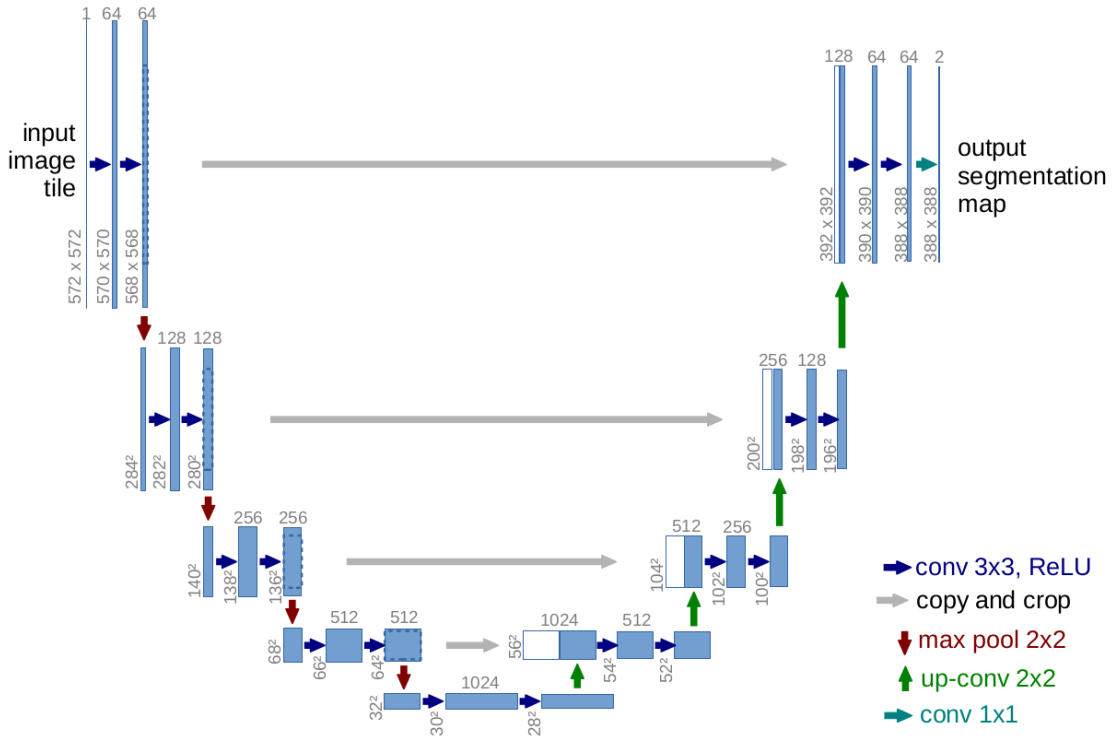


Figure 2.9: The U-Net architecture, with contractive side on the left, and expansive side on the right. The feature map follows the arrows and multiplies the number of feature maps twice at each contraction, and halves at each expansion [27].

Simultaneously, a unique architecture was under development. This architecture mirrored the hourglass structure of contracting and upsampling sections of the FCN, only this time, the illustration of its architecture led to its name, the UNet [27].

The UNet similarly consists of a contractive and expansive side, with the added feature of so-called 'skip-connections'. As shown in Figure 2.9, at each stage of the contracting/expansive side, these copy operations help localise high-resolution features from the contracting path [27]. The network yielded great results for a few training images and more precise segmentations.

The U-Net was now very close to being a staple choice in biomedical data for anatomy segmentation. A limitation of its current implementation is that slices along the third dimension would most certainly contain contextual information that would influence the decision of the current design. Therefore, an identical network topology with extended 3D convolutions was proposed by Çiçek et al. [28].

2.2.3 nnUNet

The nnUNet is closely related to the U-Net architecture. However, nnUNet's ability to adapt to the data improves this model performance over the vanilla implementation.

Until now, architectures have operated on data pre-processed to a particular expected distribution and configured to deal with a specific problem. For instance, out-of-the-box configuration has stringent input size requirements, which require image resizing. Furthermore, 3D images commonly produce heterogeneous voxel spacing depending on the parameters chosen by the clinician or the machine from which they were produced [29]. The number of moving parts often leaves a unidirectional dependency on the data depending on the architecture.

Automated Method Configuration

Therefore, the nnUNet analyses the fingerprint of the dataset and the device to deliver a tailored experience and force a more codependent relationship; now, the architecture depends on the data and the data is pre-processed to conform to the network [29]. Furthermore, hardware restrictions mean networks may be inaccessible to those with worse specifications or, at the other end of the spectrum, may underutilize powerful computation still available [29]; the nnUNet analyses GPU constraints used to influence batch sizes and more [30].

The automated method configuration is classified into three categories. A dataset fingerprint extracts training data distributions such as shape, spacing and intensity distributions. Rule-based parameters estimate the most common robust parameters for resampling and normalization. Finally, the Empirical Parameters learn parameters, such as ensemble selection, which is not derivable from the dataset fingerprint.

Critical Review of the nnUNet

A review of segmentation methods in 2024 reviewed some further developments in segmentation models and found that the convolution-based U-Net architectures continued to outperform Attention-based or Mamba-based approaches six years after the initial publication of the self-configuring network [31]. Isensee et al. concluded that there was a significant mischaracterisation of proclaimed improvements in new strategies such as transformers. Claims of performance improvements over the nnUNet were reviewed through the control of validation datasets and removals of baseline tampering, which demonstrated the convolution-based performance on datasets with low statistical intra-method standard deviation [31].

The continued performance dominance gives the nnUNet a good foundation for being used as a baseline model for all datasets.

2.2.4 TotalSegmentator

TotalSegmentator is a tool based around the nnUNet. TotalSegmentator is pre-trained on 1204 CT examinations to provide plans to segment 104 anatomical structures. The anatomies selected included apparent structures such as skeletal structures, gastrointestinal organs and other major organs. The training data contained many CT images, with differences in slice thickness, resolution, and contrast phase [32]. However, it is important to note that 60% of the scans occurred in a contrast-enhanced environment, which plays a role in how obvious delineations are during scanning, a scanning detail that was omitted during the training data

collection for this research project. Furthermore, only an estimated 10% of the data collected from this model contained relevant studies for the abdomen and pelvic areas.

From the segmented organs that total segmentator provides, only the Bladder overlapped with the organs that were of interest in this study.

2.2.5 UniverSeg

Convolutional architectures like those discussed above utilize many-shot learning (Section 2.3.1). However, models trained on segmenting a target domain (e.g., bone delineation) do not transfer to other domains (e.g., organ segmentation) without fine-tuning.

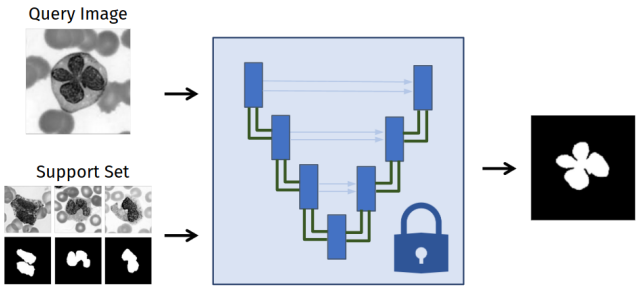


Figure 2.10: The UniverSeg architecture [33]. Diagram illustrates the freezing of model parameters while providing a support set of images which follows the query image through the segmentation pipeline.

Butoi et al. present UniverSeg, a model that breaks away from the traditional approach; instead of training a model on a single task (e.g. bone delineation) and freezing parameters during inference, this architecture uses a support set of images to provide a practical few-shot approach to inferring segmentations from input images. Figure 2.10 shows an example of this efficient querying, which manifests itself in a U-Net architecture where the support set passes through the network along the query to influence the final segmentation. This way, Butoi et al. attempt to provide segmentation on a target image based on examples of other samples with the same anatomy contoured in a selection of other images. This way, the model can learn to segment both bone and organ segmentation tasks without the need for fine-tuning [33].

This model operates on 2D slices of images and directly avoids the finetuning argument for medical imaging. They argue that finetuning can be unhelpful due to the differences between medical domains, features, and data fingerprints. As such, UniverSeg avoids significant retaining for each subtask.

2.2.6 SAM

Advances in NLP, with attention-based mechanisms, have questioned whether convolutional-based methods like those above are the best approach for segmentation. Transformers from NLP were adapted to form the Vision Transformer (ViT) [35]. This model views the image as a grid of tokens; in the original paper, Dosovitskiy et al. separate the image into a grid of patches and read in these grid cells as individual tokens. The tokens pass through the attention mechanism as with NLP and into a classification mechanism [35].

The SAM model implemented a modification of the transformer architecture [34] as seen in Figure 2.11. However, the task was reformulated as a promotable segmentation problem to

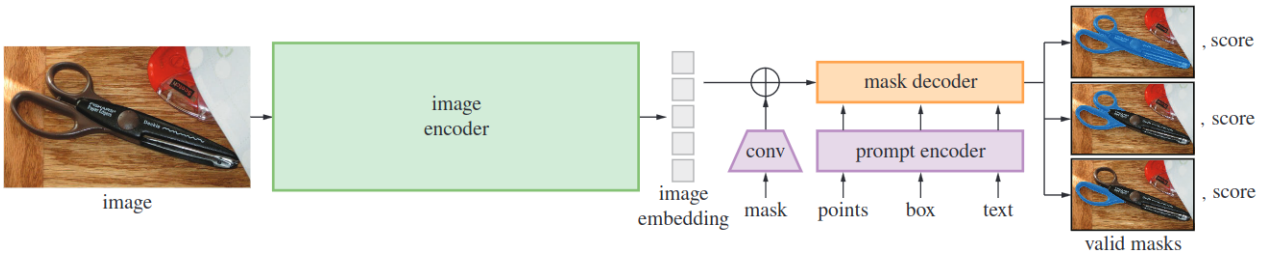


Figure 2.11: The SAM model involves a transformer architecture that embeds points and bounding boxes into a promptable encoding, which is used in tandem with the image encoding to produce the most likely segmentation of the described area [34].

allow for zero-shot generalisation (the model can generalise to unseen examples with no retraining or fine-tuning). As seen in Figure 2.11, the model inputs an image along with either points in the image or boxes. This reduces the search space SAM has to perform to segment an object into an area or a set of points.

2.2.7 MedSAM

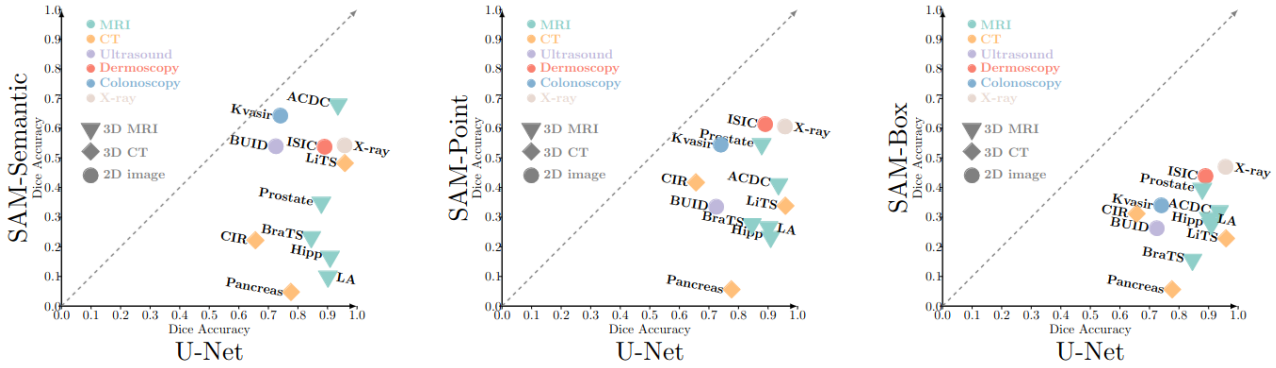


Figure 2.12: Performance of a SAM model across a reference nnUNet baseline when applied to different datasets. The evaluation was considered for semantic, point and box prompts. [36].

SAM trains its network on a collection of natural images, not medical images such as CT and MRI scans. Extensive stress tests performed on SAM concluded that SAM's out-of-the-box promotable segmentation tool had good baseline performance on large visible objects [37] but required exact prompt segmentation, making it inaccessible to automated contouring. Regarding the number of points required to make a sensible prediction, SAM quantitatively underperformed a nnUNet baseline, with qualitative evaluation showing fuzzy boundaries in medical contexts [38]. Finally, Figure 2.12 shows the performance of SAM against an nnUNet baseline across a set of datasets and imaging modalities [36] which concludes that SAM never outperformed the nnUNet baseline when trained on the 11M natural images [34].

Therefore, Ma et al. enhanced the architecture provided by SAM by training it on a dataset of nearly 500k CT scan test examples to train a model for medical images, named MedSAM [39]. Ma et al. decided to keep close to its original despite medical images being 3-dimensional in CT and MRI scans because of "enhanced flexibility and adaptability" where slices along an axis substitute 3D scans [39]. This model demonstrates an improvement over SAM, nnUNet, and Deepmedic models when MedSAM bounding boxes extracted from the ground truth prompt the model [39].

2.3 Transfer Learning

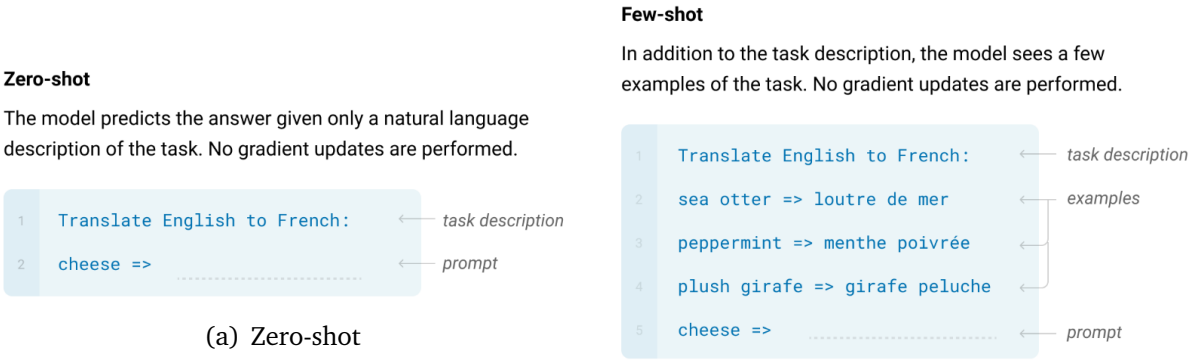
Transfer Learning involves transferring a model's knowledge from a domain that has trained on more volumes of training data in another domain and is used as a starting point in the target domain. Transfer is a source of success because, in the early layers of a model, it typically learns very low-level features. At this scale, the objective of the original domain does not matter; regardless of the initialisation, a model working on a similar problem will inevitably learn similar low-level features.

Arguably, the universality of the parameters learnt in a model with prosperous access to data will have richer and better patterns that another with less data did not have enough information to learn [40, 41].

Transfer Learning has the potential to improve initial performance using only the transferred knowledge before any further learning begins, improve the time it takes to thoroughly learn the target task given the transferred knowledge, and improve the final performance all when compared to initial benchmarks without transfer [42]. Medical contexts have already applied Transfer Learning, which reportedly improved weight initialisation for 332 abdominal liver CT scans and resulted in faster convergence, providing a more robust representation [43].

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. The prevention is because the model has already learnt features likely to be helpful in the second task [44].

2.3.1 Shot based learning



(a) Zero-shot

(b) Few-shot

Figure 2.13: Shot based learning captured in the natural language context [45]

The transfer of models is dependent on the strategy the model uses to make predictions. Shot-based learning is a substantial factor. This describes the model's exposure to data in domains and its ability to handle cases it has never seen before. Thus, shot-based learning can be separated into meta-learning tasks or multitask learning.

Instead of learning underlying patterns, meta-learning models learn to *learn* the algorithm itself. This allows it to generalize tasks with few labelled examples of new or rare cases. Otherwise, multitask learning makes predictions for a particular set of tasks [40].

Zero-shot Learning

Zero-shot models (Figure 2.13(a)) can generalize to unseen tasks. An example of a model targeted for zero-shot transferability is the SAM model (Section 2.2.6).

Few-shot Learning

Few-shot models (Figure 2.13(b)) can generalize to unseen tasks with a few examples. The UniverSeg model is the target model for few-shot transferability (Section 2.2.5).

Many-shot Learning

Many-shot models can generalize to unseen tasks with many examples. The nnUNet based models such as the TotalSegmentator is the target model for many-shot transferability (Section 2.2.3).

Chapter 3

Methodology

Improvements over the transfer from models trained on other domains could hypothetically allow the segmentation of delineated areas more accurately. To investigate this claim, we iterate over the three types of shot learning, zero-shot, few-shot, and many-shot learning, to determine the effectiveness of transfer learning in the radiotherapy domain.

3.1 Baseline – nnUNet

The provided examples of the objective are enough to train an nnUNet model from scratch to segment anatomy and radiotherapy target volumes accurately. This nnUNet model has been shown to provide more robust results in external stress tests than other comparable architectures on a set of different datasets [31]. Also, in many biomedical applications, "only very few images are required to train a network that generalises reasonably well" [28].

3.1.1 Preprocessing

Prior to training, necessary normalization must take place to standardize each input. Firstly, CT scans can produce results for different spacings, resolutions, and dimensions. Because the model samples data in batches, the batch properties must align within the model's body. Therefore, the fingerprint taken by the preprocessing pipeline resamples input data towards the median dimension.

The traditional method of normalization involves normalizing the entire image corpus. However, this method does not take into account the skew that might affect the outcome. This is because the background value occurs most frequently, and artifacts such as metal cause outlier peaks. As a result, traditional normalization could overlook important tissues as it attempts to treat background and outlier values equally with foreground values.

Therefore, the nnUNet avoids this complication by processing voxel properties encased by the ground truth segmentation. The values are clipped to their 0.5 and 99.5 percentile, followed by traditional normalization with the mean and standard deviation. This way, the target structure properties remain relative to their original, and the background label conforms to the normalization inspired by the region of interest.

3.1.2 Separate Training

The default strategy is to consider each anatomy separately and attempt to learn segmentation patterns without considering constraints mentioned in Section 2.1.8. These models can be used in an ensemble system to produce thorough segmentations of the target volumes, oblivious to clinical constraints.

This will act as the baseline capabilities of models with no transfer of knowledge.

3.2 Many-shot Transfer – TotalSegmentator

We kickstart the transfer discussion by evaluating the transfer of many-shot models. The candidate model for this is the TotalSegmentator model, which is entirely based on the default implementation of the nnUNet.

When transferring information, many-shot models are the most obvious choice. These models have the potential to increase the amount of helpful information that can be used to segment anatomies in the target domain by transferring the original task. The hypothesis is that many-shot transfer models will improve segmentations for clear organ delineations. Additionally, anatomies that have already been segmented (such as the bladder) will fine-tune to become more accurate.

3.2.1 Separate Training

Like the nnUNet, we apply the default fine-tuning strategy to a pre-trained TotalSegmentator model. TotalSegmentator has 23 separate nnUNet pre-trained models on different sub-tasks in the anatomical segmentation [46]. These individual models segment structures like the vertebrae, cardiac muscles, rubs, and lung vessels. For this destination delineation task, we chose the 'organ' model trained on a withheld dataset of 1200 CT scans. The only anatomy shared between the two domains is the Bladder which will hypothetically provide a performance improvement for this anatomy due to the additional thousand examples of the bladder.

3.2.2 Region Based Training

The natural extension of separate training is to consider training each class simultaneously. Region-based training is a noninvasive method offered by the nnUNet. This training style combines the 3D segmentation maps for each of the seven classes into one 3D segmentation by introducing new IDs wherever there is a new overlap. An example slice can be seen in Figure 3.1.

For the purposes of evaluation, the following methods will be applied to the baseline nnUNet model to assess the success.

No Rule Enforcement

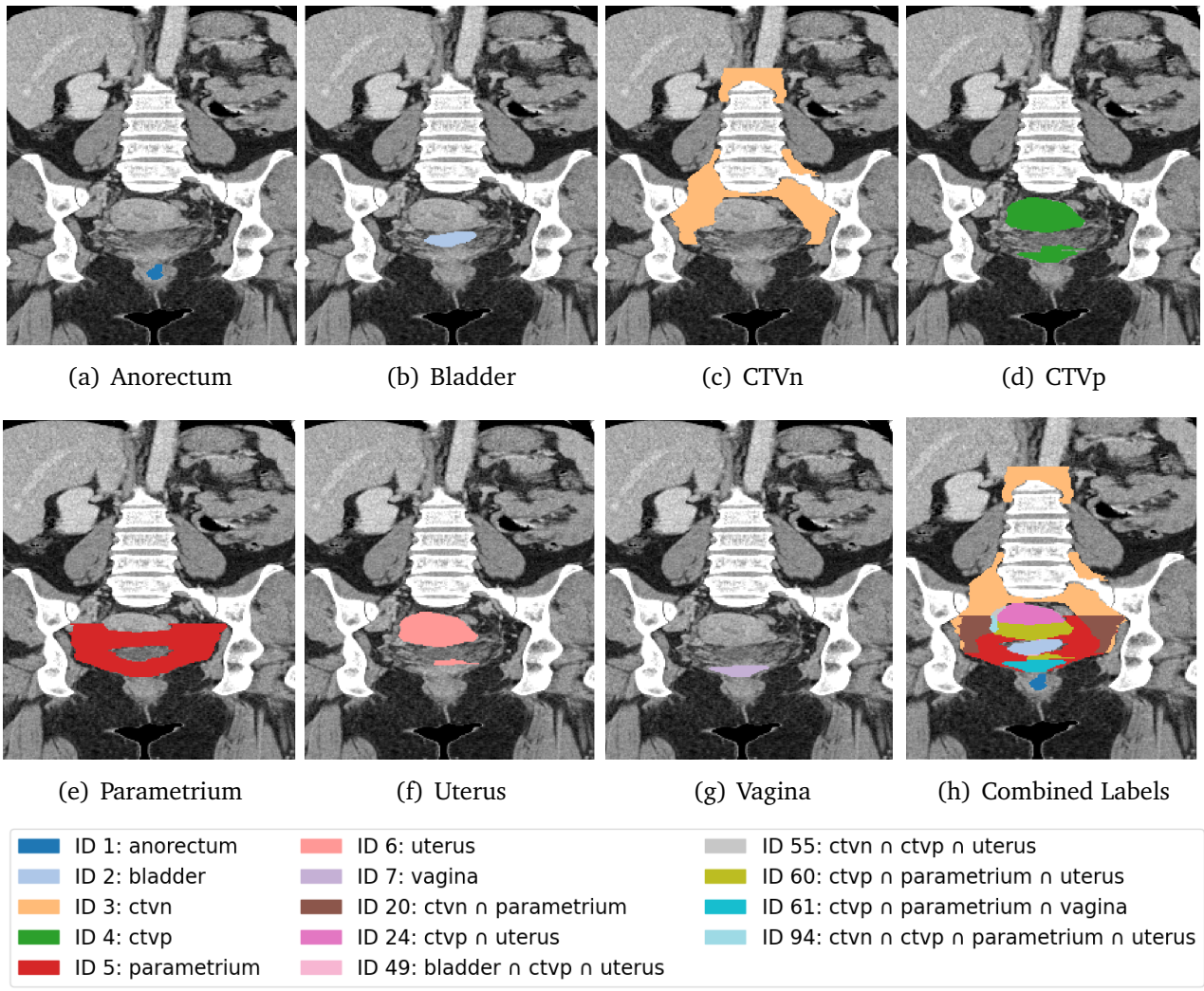
A sensible and calibrated set of ids could be drawn based on the rules in Section 2.1.8. However, practical experience and the recalcitrant and challenging nature of defining logical expressions to generalize something as complex as organs and something as fuzzy as microscopic cancer spread mean that, in practice, these rules cannot be implemented strictly.

Please refer to Equation 2.1 “There should be no overlap between the CTVn, CTVp or Anorectum”. We see marginal overlap between the structures in Figure 3.1(h) in ids 55, 94. These are almost not visible on the figure, but captured by the legend.

With Rule Enforcement

We can conclude that these rules are not strictly followed in practice but should be implemented for generalisation. Therefore, an augmentation of the nnUNet trainer, specifically the [TODO], is necessary to capture this result.

We hypothesise that although the baseline model attempts to approximate the ground truth, which may call for marginal overlap, with the rules in place, this may generalise better over other cases.



(a) Anorectum (b) Bladder (c) CTVn (d) CTVp

(e) Parametrium (f) Uterus (g) Vagina (h) Combined Labels

Figure 3.1: New IDs generated for a single slice as a consequence of practical experience show non-conformity to rules in Section 2.1.8.

3.3 Few-shot Transfer – UniverSeg

UniverSeg operates on 2-dimensional data. As a result, to provide automatic radiotherapy planning volumes, it is essential to provide slices for the model as input alongside a sufficient

support size of similar slices.

3.3.1 Preprocessing

Images are normalized to match UniverSeg's training properties. Specifically image intensities are clipped to the range $[-500, 1000]$ and normalize to be between $[0, 1]$ and finally scaled down to a 128×128 resolution [33].

For the purpose of the experiment, this was repeated along each axis to produce a

3.3.2 Selecting support

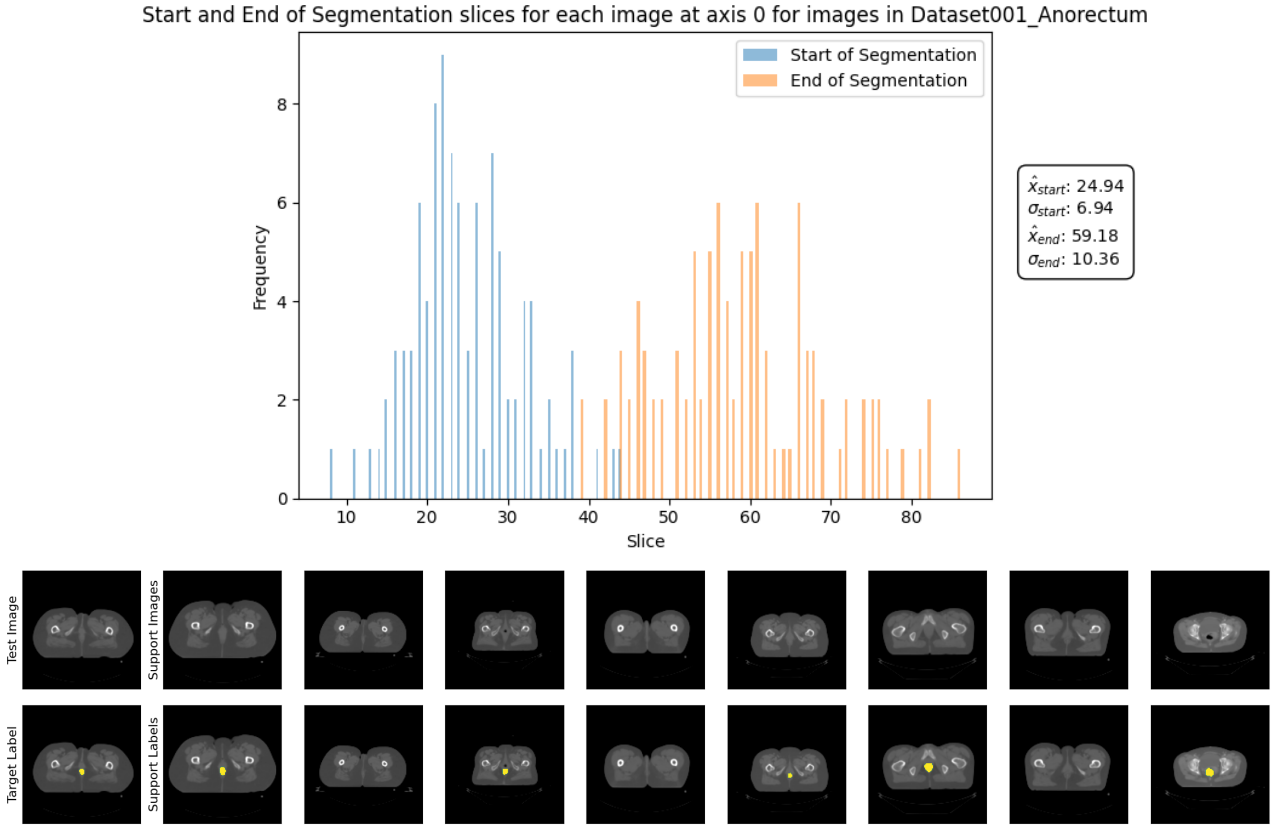


Figure 3.2: The distribution of the start and ending slices of the tumor in a normalized batch of images across axes 0 (axial dimension) and the corresponding sampled support extracted.

However, for an unseen example, the location of the tumour can only be determined by referring to assisting models that provide estimates. In a secluded environment where only one's model transferability is assessed, we instead analyse the properties of the tumour from the examples.

We begin by analysing the tumour locations along each axis. The intuition is that the normalised images of equal dimensions and spacings should contain the tumour at approximately the same location. Figure 3.2 shows the distribution of start and end slices for the tumor and the corresponding sampled support set.

3.4 Zero-shot Transfer – MedSAM

MedSAM is the final algorithm in the shot-learning strategies that will be covered. Specifically, MedSAM is being evaluated on the provided dataset, both with and without transfer, to test the claim of transferability.

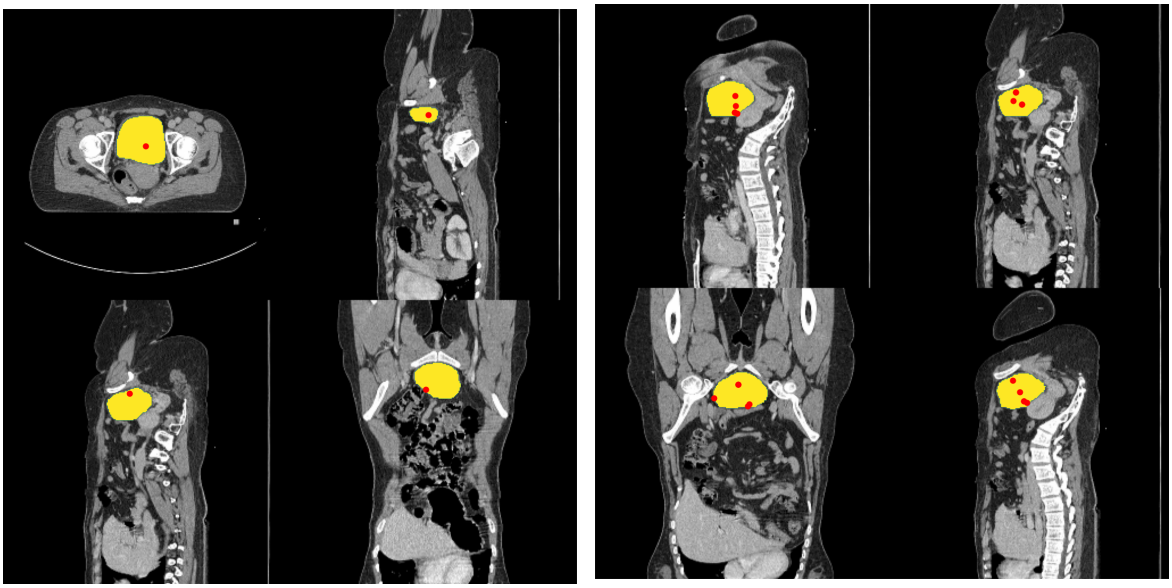
Zero-shot learning, a promising feature of MedSAM, could potentially enable it to handle tasks it has never encountered before. However, the assertion that it can seamlessly transfer without refinements is a bit of a stretch. This is because certain contours, like the CTVn, delineate tiny tumor spread throughout the lymph node system. Therefore, it is unlikely that MedSAM, trained in such cases, would be used for these specific tasks. It would typically be used for more straightforward volume delineations, such as organs.

3.4.1 Preprocessing

MedSAM takes two-dimensional images as input, much like UniverSeg. Therefore, we similarly preprocess the data to a resolution of 1024×1024 along each slice where the ground truth is present. We clip the ranges of the CT scan to Hounsfield units centred around the 40 value, with a window radius of 400 units and later normalized to the range of $[0, 1]$.

3.4.2 Point based transfer

We first experimented with the MedSAM architecture to test the effectiveness of prompting on points. We formulated this task into two categories: a sparse and dense point experiment.



(a) Sparse point prompt (b) Dense point prompt

Figure 3.3: The two strategies for sampling structures in the MedSAM model.

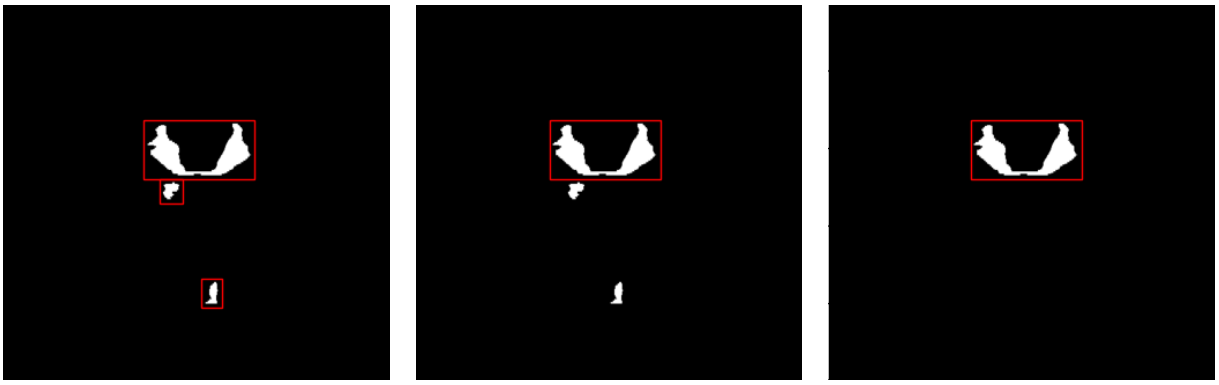
The two strategies are shown in Figure 3.3. The sparse point prompt is a single point in the image, while the dense point prompt is a set of n points randomly sampled from the ground truth segmentation.

The hypothesis is that point prompts may work well with structures such as the bladder which are trivial to identify. However, for the CTV, the point prompt may not segment the full extent of the structure.

3.4.3 Box based transfer

A box prompted solution will define the area where the tumour is likely to be. Box prompts for training are obtained from the ground truth, with a margin surrounding the box.

For training, only one box may be fed into the model for inference. Therefore, we select a random structure from the boxes encompassing the tumour for a given slice. Figure 3.4 demonstrates the tumour box selection at training time.



(a) Boxed ground truth image (b) Selected tumor boxed (c) Ground truth augmentation

Figure 3.4: The process of selection bounding box for a given boxed tumor with an augmented ground truth used at training time for the MedSAM bax-based prompt model.

3.5 Quantiative Evaluation of Segmentation

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 [47].

3.5.1 Classification Based

Assesses if voxels within and outside the auto-contour have been correctly labelled [47]. 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 [48]) measures the proportion of successfully correct predictions. The Recall (also known as True Positive Rate [48]), 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".

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

3.5.2 Spatial Overlap Based

Similarly to classification-based metrics in Section 3.5.1, an overlap-based metric measures the extent of overlap between an auto-contour and a reference structure [47].

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 [47, 48, 49].

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

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 [50].

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.

3.5.3 Surface Based

Also commonly known as Boundary-Distance-Based Methods [51] compares the distance between two structure surfaces. These can be maximum, average or distance at a set percentile of ordered distances [48].

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 [51]. Therefore, the better the HD metric, the smaller the value it returns. Here, the distance is 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; therefore, direct HD application gives uninspiring results because noise and outliers are common in medical segmentations [51]. Therefore, we can calculate the average directed Haussdorf Distance.

3.5.4 Volume Based

Volume-based metrics consider only the volume of the segmentation [52, 47, 51]. However, its poor spatial descriptions make it more commonly used jointly with other metrics.

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

3.5.5 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.

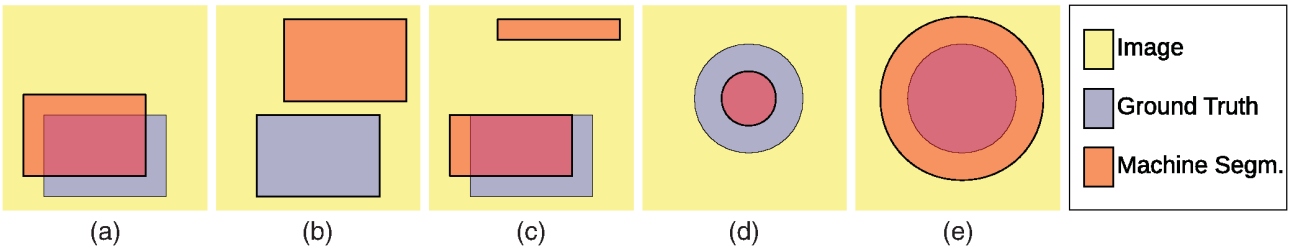


Figure 3.5: Figure from [51] illustrating cases of segmentation to aid with explanation of setbacks of certain evaluation metrics

- Classification Based (Section 3.5.1) and Spatial Overlap Based (Section 3.5.2) are similar; they are concerned with the number of correctly classified or misclassified voxels without taking into account their spatial distribution. Here, Figure 3.5(a) and Figure 3.5(c) would achieve similar results despite Figure 3.5(a) being locally bound to a better area.
- With Haussdorf Distance (Section 3.5.3) output segmentations generated by Figure 3.5(d) and Figure 3.5(e) will result in the same score, which is not favourable in a radiotherapy planning environment where an organ-at-risk is involved.

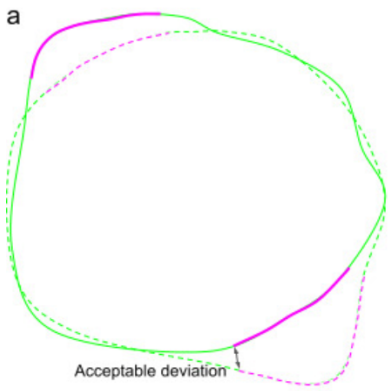


Figure 3.6: Taken from [53]. Illustrates the computation of the surface DICE, where the continuous line is the predicted surface, and the dashed line is the ground truth. The black arrows show the maximum deviation tolerated without penalty; therefore, in pink are the unacceptable deviations and green otherwise.

- Figure 3.5(b) would score flawlessly when using volumetric score estimation. However, it does not consider spatial placement, making this measurement poor when used individually.

3.5.6 Estimated Editing Based

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" [49].

Surface DSC

The study at [49] helped drive an initiative to combine aspects of surface Based evaluation (Section 3.5.3) and Spatial Overlap Based evaluation (Section 3.5.2) into a Surface DICE which assesses the specified tolerance instead of the overlap of the two volumes.

We can formulate the Surface DSC score in a mathematical definition [49] with its corresponding illustration in Figure 3.6.

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

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" [50]. 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 3.5.6)

3.5.7 Summary

This is why we settle at the Surface DSC (Section 3.5.6), 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 3.5(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 3.5(d) instead of Figure 3.5(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 [53].

Chapter 4

Results and Discussion

Chapter 5

Conclusion

Transfer works!

Chapter 6

Ethics

The lack of effort to protect the identities and confidentiality of patients during research projects may result in “stigma, embarrassment, and discrimination” [54] 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.

6.1 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. [55]. Imperial College’s Medical Imaging team also arranges formalities, such as acting as ”ethical data stewards” [56].

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 [57]. 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.

6.2 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 [58].

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 [59].

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 [60].

Clinicians can fall into the trap of automation bias as AI becomes more commonplace in clinical environments [61]. 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 11

[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 11

[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 11

[25] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation, 2015. pages 11, 12

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

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL `http://arxiv.org/abs/1505.04597`. pages 12

[28] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016. URL `http://arxiv.org/abs/1606.06650`. pages 12, 18

[29] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Köhler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation. 2018. URL `https://arxiv.org/pdf/1809.10486.pdf`. pages 13

[30] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. 2021. URL `https://www.nature.com/articles/s41592-020-01008-z`. pages 13

[31] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation, 2024. pages 13, 18

[32] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. 2023. URL `https://arxiv.org/pdf/2208.05868.pdf`. pages 13

[33] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation. 2023. URL `https://arxiv.org/pdf/2304.06131.pdf`. pages 14, 21

[34] Alexander Kirillov, Eric Mintun, Nikila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. Technical report, Meta AI Research, 2023. URL https://arxiv.org/abs/2304.02643. pages 14, 15

[35] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929. pages 14

[36] Sheng He, Rina Bao, Jingpeng Li, Jeffrey Stout, Atle Bjornerud, P. Ellen Grant, and Yangming Ou. Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets, 2023. pages 15

[37] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W. Remedios, Shunxing Bao, Bennett A. Landman, Lee E. Wheless, Lori A. Coburn, Keith T. Wilson, Yaohong Wang, Shilin Zhao, Agnes B. Fogo, Haichun Yang, Yucheng Tang, and Yuankai Huo. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging, 2023. pages 15

[38] Chuanfei Hu, Tianyi Xia, Shenghong Ju, and Xinde Li. When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation, 2023. pages 15

[39] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, Jan 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z. URL https://doi.org/10.1038/s41467-024-44824-z. pages 15

[40] Christopher M. Bishop and Hugh Bishop. *Deep Learning, Foundations and Concepts*. Springer, 2023. pages 16

[41] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. 2009. URL https://ieeexplore.ieee.org/abstract/document/5288526. pages 16

[42] Lisa Torrey and Jude Shavlik. Transfer learning. Technical report, University of Wisconsi, 2009. URL https://ftp.cs.wisc.edu/machine-learning/shavlik-group/torrey.handbook09.pdf. pages 16

[43] Michal Heker and Hayit Greenspan. Joint liver lesion segmentation and classification via transfer learning. Technical report, 2020. URL https://arxiv.org/pdf/2004.12352.pdf. pages 16

[44] What is transfer learning? URL https://www.geeksforgeeks.org/ml-introduction-to-transfer-learning/. pages 16

[45] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models

are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165. pages 16

[46] wasserth, lassoan, fedorov, cnicolasgr, and Arputikos. URL https://github.com/wasserth/TotalSegmentator. pages 19

[47] 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 23, 24, 25

[48] 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 24, 25

[49] 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 24, 26

[50] 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 24, 26

[51] 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 25

[52] 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 25

[53] 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 26, 27

[54] 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 30

[55] 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 30

[56] 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 30

[57] *Data Protection Act 2018*. URL https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted. pages 30

[58] 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 30

[59] 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 30

[60] 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 31

[61] 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 31