Imperial College London

MEng Individual Project

Department of Computing

Imperial College of Science, Technology and Medicine

[INTERIM]: Transfer Learning for Deep Learning Radiotherapy Planning

Author:
Anton Zhitomirsky

Supervisor:
Prof Ben Glocker

Second Marker:
Dr Thomas Heinis

May 23, 2024

Abstract

Cervical cancer remains as one of the top cancerous diseases to affect women. To treat it, Oncologists plan a contour for therapy after obtaining 3D contrasting images of soft-tissue organs at risk and tumorous areas.

Auto-segmentation differs from Auto-contouring tasks due to lacking clinical knowledge surrounding the location of the cancer and biological spreading patterns. Instead of trivially contouring visible macroscopic tumour masses on a scanned patient, a clinician requires also to adjust for microscopic spreads and finally error margin spreads. This target volume should aim to treat the disease in one-shot and not affect any organs-at-risk.

The scientific community has tried to automate this task using current architectural standards such as CNNs or U-Net based algorithms. However, no studies yet consider Transfer Learning as an approach to solving this issue. This report investigates this architectural challenge to contribute to the total pool of deep learning auto-segmentation models in a communal effort to save resources of medical institutions.

You can find the most up-to-date version of the report *here*.

Contents

Chapter 1

Introduction

1.1 Clinical Context

In 2017, Cervical Cancer accumulated 530,000 new cases annually, with 270,000 deaths, making it the fourth most common malignancy diagnosed in women worldwide [**cervical-cancer-epidemic**]. A common treatment mechanism involves radiation therapy which targets cancerous cells in a clinically defined target area with beams of high energy (Section 2.1.2). This treatment is tedious, as it is estimated that an oncologist needs 90-120 min to delineate target areas for radiotherapy [**LIU2020184**].

Radiotherapy has become a great option due to high resolution X-ray or CT scans which produce high contrasting images of the damaged and surrounding soft tissue [**radiotherapy-basic-concepts**]. Physicians then use this 3D scan to plan a target volume for the radiation therapy surrounding the tumor in hopes of killing it and not damaging the surrounding tissue.

Areas are therefore constructed based on an Oncologist knowledge about the particular cancer to determine target structures, structures we need to protect (organs-at-risk), and areas where each particular cancer is likely to spread to [**AMLART-data**]. These areas are delineated onto a patient scan and used as a radiotherapy target used for treatment.

Accurate scans have been provided to see if an AI model can learn cervical cancer CTV patterns adjacent to clinical knowledge and oncologist prior knowledge. Training models which produce substructures required for radiotherapy target volumes would overall save time and improve consistency within the radiotherapy planning process [**AMLART-data**].

1.2 Current Solutions

The problem of automatic delineation of tumours in a patient is not a new concept unfamiliar to the scientific community. From 2016 to 2020 the number of deep-learning in radiotherapy publications has grown from 1,001 to 3,653 [**Lin2021-oz**]. Therefore, there exist many proposals and solutions to this problem for cancerous tumours across the body. This also includes research projects focused specifically on cervical cancer and radiotherapy auto-contour planning. By conducting a accrual of literature across PubMed with the search string

> "radiotherapy" AND "contour" AND ("cervix" OR "cervical") AND "cancer" AND ("Deep Learning" OR "DeepLearning" OR "Machine Learning" OR "ML" OR "Artificial Intelligence" OR "AI" OR "Computer assisted")

6 key papers published between the span of 2020 to 2021 were selected as most relevant for this project [**Samarasinghe2021-ps**, **Lin2021-oz**, **Sartor2020-et**, **LIU2020184**, **Rhee2020-ms**, **LIU2020172**]. These papers proposed novel ideas on how to solve image segmentation problems. Some noteworthy networks included Convolutional Neural Networks (Section 2.2.1), 2D and 3D U-Net adaptations (Section 2.2.2 and 2.2.3), V-Nets and DeepMedic models, with the most common approach being the U-Net approach according to a clinical survey on the topic in 2021 [**Samarasinghe2021-ps**]. The

methods in these papers will be further discussed in Chapter 2.

1.3 Motivation

Healthcare Crisis

Healthcare centres are experiencing high demand with low availability for all types of cancer. In England, waiting times are getting worse each month, with over a third of cancer patients (60,000) waiting beyond the 62-day target, and 10,000 patients waiting over 104 days according to a radiotherapy manifesto in 2022 *Manifesto link (shall i put in bibliography?)*.

This wait is partially due to the time consuming nature of treatment plans, with manual segmentation performed by a professional oncologist taking 90-120 minutes [**LIU2020184**, **Sartor2020-et**] while always requiring optimization of radiotherapy placement to avoid oragns-at-risk [**Samarasinghe2021-ps**].

Furthermore, more resources can be partitioned into other departments by developing software to aid with cancer treatments. A study found that the scale-up of radiotherapy capacity in 2015-35 from current levels 'could lead to saving of 26.9 million life-years in low-income and middle-income countries over the lifetime of the patients who received treatment' [**expanding-global-access-to-radiotherapy**].

Current Tools

Unfortunately, there is a difference between auto-contouring and auto-segmentation. For our case we cannot use models such as TotalSegmentator (Section 2.3.1). TotalSegmentator has learnt form many samples to delineate 117 classes of objects in the human body, such as bones, a large subset of significant organs and veins [**totalsegmentor-git**]. Our case involves delineating a PTV, which is a significant challenge because it involves meta-information which is not visible on a CT scan, and therefore is not a trivial task of tracing around the visible contours of a tumour, but more so about adding sufficient padding to account for the microscopic spread of the tumour not visible to the scanner.

Since wrong or inaccurate contours constitute the highest factors for failure of treatment [**Rhee2020-ms**] current tools that are not bespoke for the task are therefore poor candidates for practical use.

Research Gap

The current literature makes great use of modern concepts stemming from the discovery of CNN models for auto-segmentation for radiotherapy planning. However, all papers assume a good sample size which represents the overall population of people under the same conditions. This is not practical in a general case, for instance: hospitals may only have the data for their in-house patients, teams operate on a niche areas and therefore have a smaller sample size, affected area scans are not standardized and vary between organizations which may promote the 'garbage in, garbage out' philosophy, and finally varying degrees of catastrophe levels pertaining to the sensitivity of the organs at risk (Section 2.1.3) surrounding the tumour (radiotherapy in the pelvis vs radiotherapy in the brain).

Summary

We therefore propose a solution that can be applied to groups with specific niche rules and small sample sizes. This Transfer Learning approach (Section 2.5) aims to leverage other trained models with great performance and lift low-level features to use in our case. This will also fill the literature gap which hasn't considered solving the auto-contouring for radiotherapy planning volumes, as a PubMed search querying publications with mentions of transfer learning returned no relevant matches.

1.4 Outline of Report

> ☞ This interim report will not contain implementation details yet.

The structure of this report will take the reader of a reasonable programming background through the project such that they might be able to reconstruct the outcome themselves. It is expected of the reader to understand the concept of Machine Learning and an intuition surrounding Computer Vision.

Firstly Chapter 2 discusses the clinical context (Section 2.1) from which the idea for this project originates. We further construct the background knowledge for the current Computer Vision methodologies which exist to solve segmentation based issues (Section 2.2). This will lead into currently pre-trained models (Section 2.3) and their application in a Transfer Learning setting (Section 2.5).

With collaboration from the Royal Marsden Hospital they have provided this project with a set of data (Chapter 3) with which the remainder of this project is trained upon, the usage concerns have been evaluated through the Ethics chapter (Chapter 5). We discuss what constitutes an accurate segmentation and reason about evaluation metrics to provide an outlook on the models performance (Chapter 4).

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

Chapter 2

Background

2.1 Clinical Context

This section will provide a baseline understanding of the clinical context for developing a tool to help segment tumours in patients with cervical cancer. Considering the unfortunate high frequency of cancers developing in people across the globe [**cervical-cancer-epidemic**], the idea for developing a model to help segment cancers has long been in the scope for many researchers.

2.1.1 Cervical Cancer

In 2017, Cervical Cancer accumulated 530,000 new cases annually, with 270,000 deaths, making it the fourth most common malignancy diagnosed in women worldwide [**cervical-cancer-epidemic**]. A common treatment mechanism involves radiation therapy which targets cancerous cells in a clinically defined target area with beams of high energy (Section 2.1.2). This treatment is tedious, as it is estimated that an oncologist needs 90-120 min to delineate target areas for radiotherapy [**LIU2020184**].

This time consuming nature of studying each patient makes treating cancers a time consuming endeavour which is particularly threatening when time is not on the patients side. Death from Cervical Cancer involves significant pain and suffering for the patients who cannot receive urgent treatment [**cervical-cancer-epidemic**]. This is particularly a problem in mid-low income countries with approximately 85% of cases of cervical cancer occurring, making them have 18 times the death rates of high-income countries [**cervical-cancer-epidemic**].

2.1.2 Radiotherapy

Radiation therapy is a treatment option for cancer treatment. In 2012, approximately 50% of all cancer patients received radiation therapy, with an additional 40% involving curative treatment [**radiotherapy-advances**]

Physicians abuse the physical properties of radiation to damage the genetic material of cells and block their ability to cause further damage to a patient [**radiotherapy-advances**]. Importantly for our cause, radiation damages intercellular molecules leading to degradation and stopping a cells ability to divide, leading to interphase death. Alternatively, radiation may cause a “mitotic catastrophe” causing a cell to die of a proliferative death [**cell-death**].

2.1.3 Contour Planning

Radiotherapy has become a great option due to high resolution X-ray or CT scans which produce high contrasting images of the damaged and surrounding soft tissue [**radiotherapy-basic-concepts**]. Physicians then use this 3D scan to plan a target volume for the radiation therapy surrounding the tumor in hopes of killing it and not damaging the surrounding tissue.

Process

This begins with the high contrasting image of the tumour area. In a clinical setting, an Oncologist will use knowledge about the particular cancer to determine target structures, structures we need to protect (organs-at-risk), and areas where each particular cancer is likely to spread to [**AMLART-data**].

The first area defined from the scan is the Gross Target Volume (GTV). This macroscopic delineated area is the visible tumour area on the scan and contains a high probability of containing the tumour. Secondly, the Clinical Target Volume (CTV) is derived to account for potential microscopic spread. This will be an area at least as big as the tumor area segmented in the GTV with an optional margin surrounding it containing a 'rind' of non-zero probability of tumour spread. Lastly, the Primary Tumour Volume (PTV) contains residual geometric uncertainties and safety margins surrounding the CTV ensuring the radiotherapy dose is actually delivered to the CTV [**tumor-delineation**, **defining-target-volumes**, **Lin2021-oz**, **personalised-PTV-strategies**]. These target volumes also constantly consider critical normal tissue structures which need to be preserved during irradiation. These are referred to as organs-at-risk (ORs). In some specific circumstances, it is necessary to add a margin analogous to the PTV margin around an OR to ensure that the organ cannot receive a higher-than-safe dose; this gives a planning organ at risk volume [**defining-target-volumes**].

A tangent on Accuracy

The CTV volumes may vary between clinicians as there is no internationally agreed guidelines. As a very time consuming process which high variability it therefore suffers from a lot of inter and intra-observer variability [**Lin2021-oz**]. However, the data provided has been standardized as a gold standard, see Section 3.

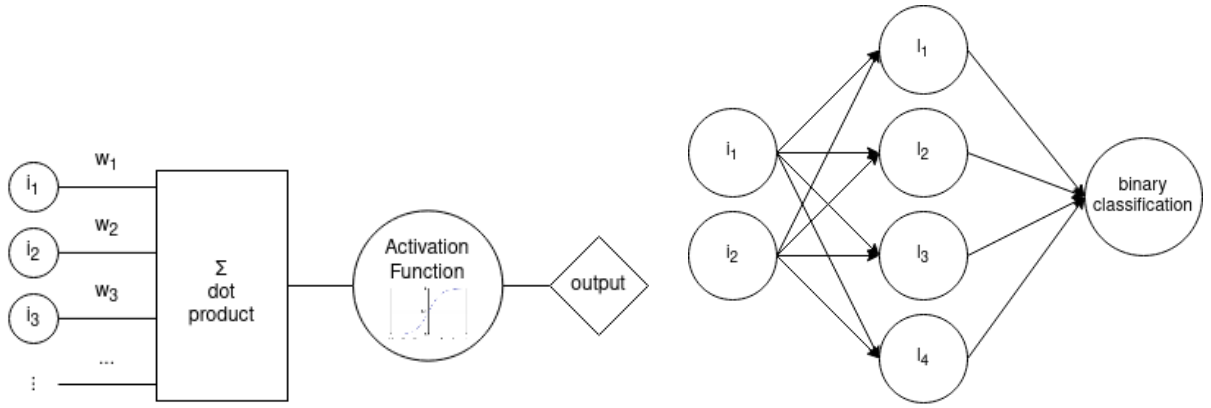
The PTV volume is an additional step which provides a margin of protection surrounding the CTV. One might think that a CTV step is the final step in clinical trials as it defines the most tight volume surrounding the cancerous site, however, geometric errors are impossible to eliminate. This includes short-term organ movement, voxel size and slice resolution, and possibility of relative movement of structure of reference and the tumour [**VANHERK200452**]

2.2 Vanilla Image Segmentation Models

Early deep learning models began by defining neurons. These mimicked biological neurons in the brain specifically by producing a set of outputs given a set of inputs. Neurons were then connected with each other to form connections. The components of this neuron would contain a set of inputs, an algebraic transformation on the inputs, an optional bias term, an activation function and then a set of outputs. By connecting individual neurons like in Figure 2.1(a) we create a connected neural network in Figure 2.1(b) to mimic the structure of the brain.

A supervised network is a network which works with labelled data. In a supervised pet classification task, an image of a cat would come with a corresponding label of 'cat'. Training the network would involve iterating on the data-set and tweaking the weights of the connections between the neurons in order to get closer to the intended result. This result is achieved by first doing a forward pass on a batch of data, obtaining the result and calculating a score. In the backpropagation step the weights are tweaked in the direction that would improve the overall performance of the network (by observing the derivative of the score for the current batch) by a certain amount (dictated by a learning rate).

The architecture in Figure 2.1(b) can be expanded to contain within it any type of complexity, both in width (height of neuron stack in Figure 2.1(b)) and depth (number of vertical neuron stacks in Figure 2.1(b)) at the cost of complexity. For instance, it is rational to think that given an input image, we could assign each pixel in the image to a neuron, and in this way, construct a fully connected network to learn patterns in the image. However, this is very computationally expensive, given that



(a) Basic building block of a neural network: The Neuron. Here, the notation i_k denotes input k, w_k denotes the weight associated with input k.

(b) A neural network comprised of individaul neurons like in Figure 2.1(a)

Figure 2.1

the number of parameters will be astronomical, and to update them all would be an intractable and inefficient solution.

2.2.1 Convolutional Neural Networks (CNN)

Work in progress. Some of my research can be seen at *here*.

2.2.2 U-Net

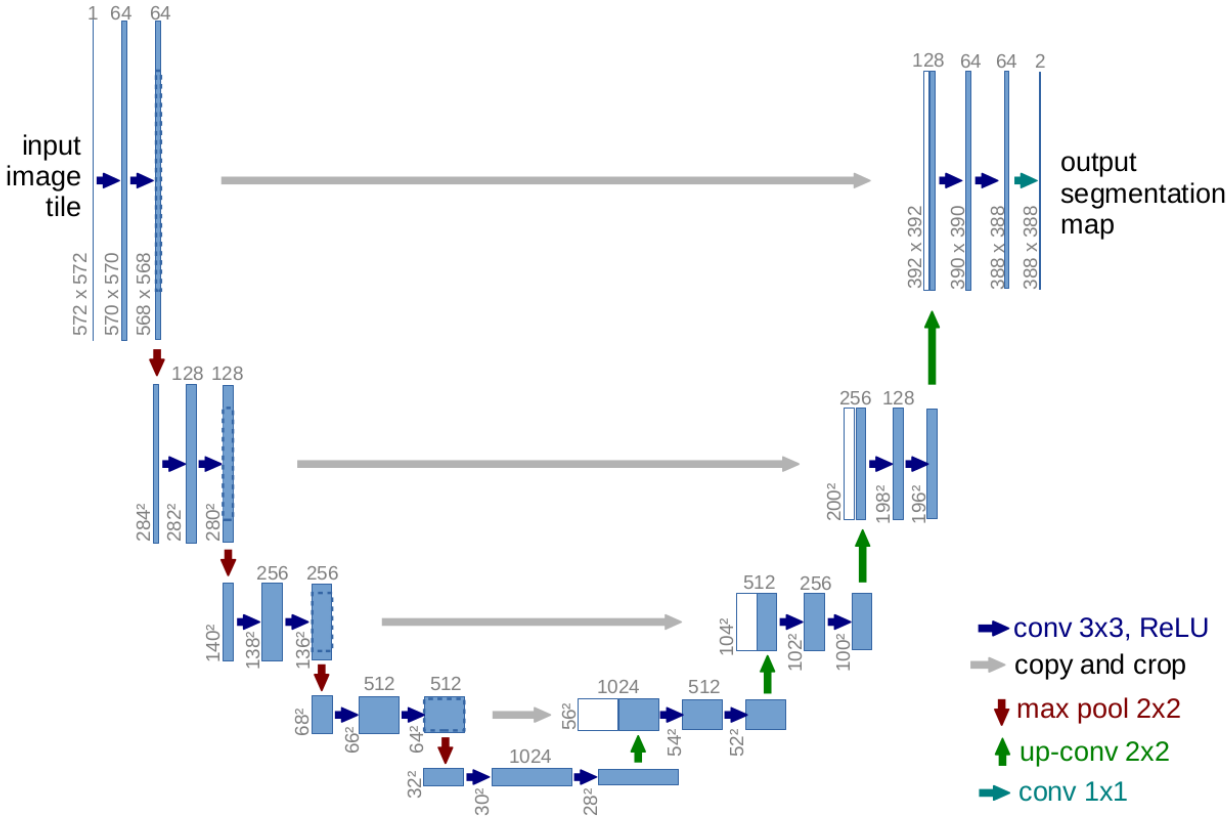


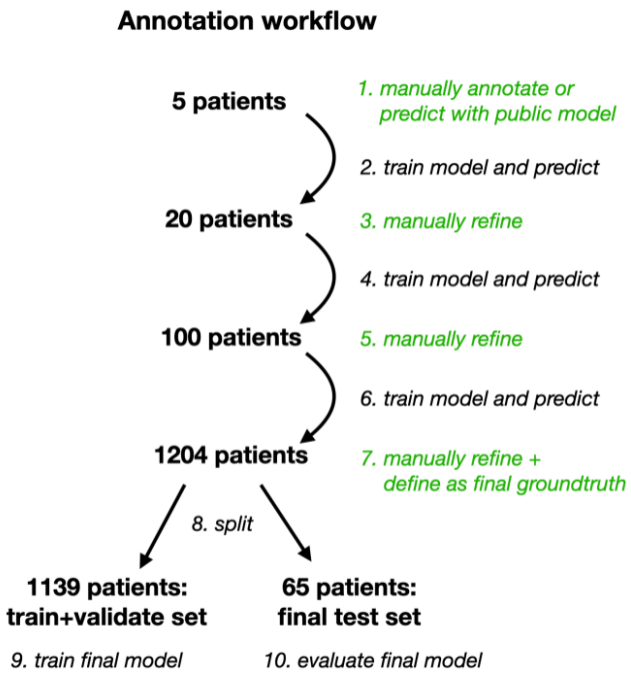
Figure 2.2: Figure of the U-Net architecture taken from [**U-Net**].

2.2.3 nnU-Net

2.3 Existing Auto-Segmentation Methods

2.3.1 Total Segmentator

TotalSegmentator [**totalsegmentor-paper**] is a model that was developed be a very robust segmentation tool. It has been trained on 1204 CT examinations and once more on 4004 whole-body CT examinations to investigate age dependent volume and attenuation changes. TotalSegmentator has learnt a robust model for delineating 117 classes of objects in the human body, such as bones, a large subset of significant organs and veins [**totalsegmentor-git**]



It uses an iterative learning approach.

(1) After manual segmentation of the first 5 patients was completed, (2) a preliminary nnU-Net was trained, (3) and its predictions were manually refined, if necessary. (4) Retraining of the nnU-Net was performed after (5) reviewing and refining 5 patients, 20 patients, and (6) 100 patients.

In the end, all 1204 CT examinations had annotations that were manually reviewed and corrected whenever necessary. These final annotations served as the ground truth for training and testing. The model was trained on the dataset of 1082 patients, validated on the dataset of 57 patients and tested on the dataset of 65 patients. This final model was independent of the intermediate models trained during the annotation workflow, which reduced bias in the test set to a minimum. Using completely manual annotations in the test set would have introduced a distribution shift and thus greater bias [**totalsegmentor-paper**].

The model uses an nnU-Net (Section 2.2.3) because of its ability yo automatically configure hyperparameters based on the dataset characteristics.

TotalSegmentator is a model which encountered some issues with the dataset that may have impeded its performance. Firstly, some patients had ribs missing, which a clinician would typically count from top/bottom to identify them; however, on some scans these weren't visible, much like in our case, we are missing overies. These have been attributed to reasons for low performance for these structures, which warns us; our dataset contains many abnormalities according to the descriptions of each patient (Section 3.3).

2.3.2 UniverSeg

2.3.3 SAM

TODO: maybe replace with medical SAM paper.

2.4 Current Limitations

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

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

2.4.1 Data Size

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

2.4.2 Bespoke Application

Another issue lies in the bespoke nature of this application. Most pre-trained networks currently run segmentation on structures are that are more obvious in a given image modality. For instance, TotalSegmentator has learnt a robust model for delineating 117 classes of objects in the human body, such as bones, a large subset of significant organs and veins [**totalsegmentor-git**]. Our application is unique because the PTV often includes a margin surrounding the visible tumour on the scan, which is different to other approaches which outline the boundaries of structures.

2.5 Transfer Learning

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

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

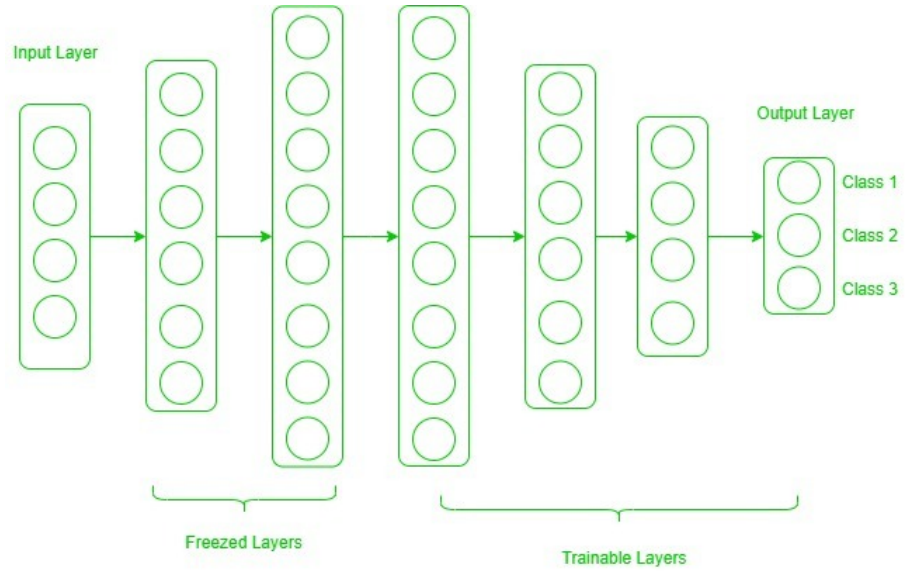


Figure 2.3: Early layers learn low-level features for similar domains, and during transfer of knowledge, these layers are frozen and the trainable layers are appended and weights are only updated for this layer [**geeks-transfer-learning**]

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

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

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

Chapter 3

Data

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

Delineated labelled data have been labelled consistently by the provider to improve chances that an AI model can learn cervical cancer CTV patterns [**AMLART-data**].

3.1 CT scan

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

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

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

3.2 File Format

The files are stored in a `.nii` file format which defines a style of image called the 'Neuroimaging Informatics Technology Initiative' (NIfTI) [**file-formats**]. It serves as a lightweight alternative to other formats such as DICOM and eliminates ambiguity from spatial orientation information [**dicom-to-nifti-conversion**].

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

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

Table 3.1: Description of NIfTI header parameters relevant to this project [**dicom-to-nifti-conversion**, **nifti-headers**, **nifti-data-format**]. Example values are taken from patient id:075.

3.3 Notes

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

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

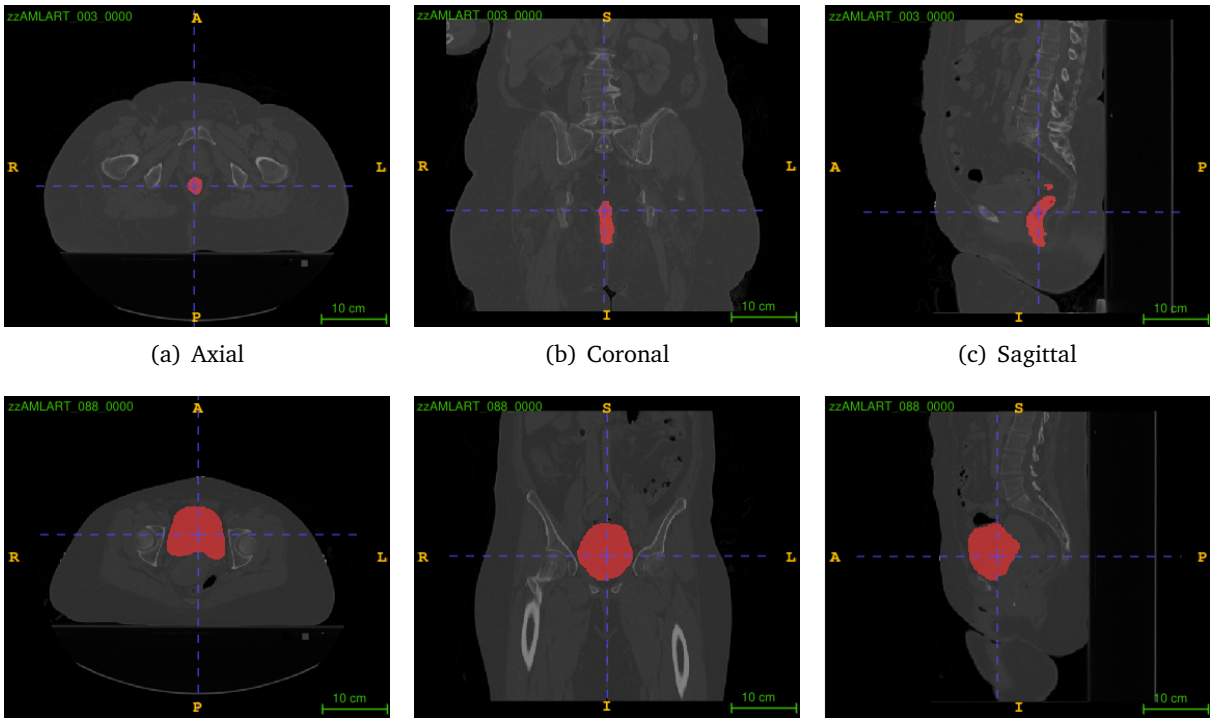
Table 3.2: A few captivating notes about each patient and why it might be concerning

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

3.4 Delineation classes

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

3.4.1 Organs At Risk



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

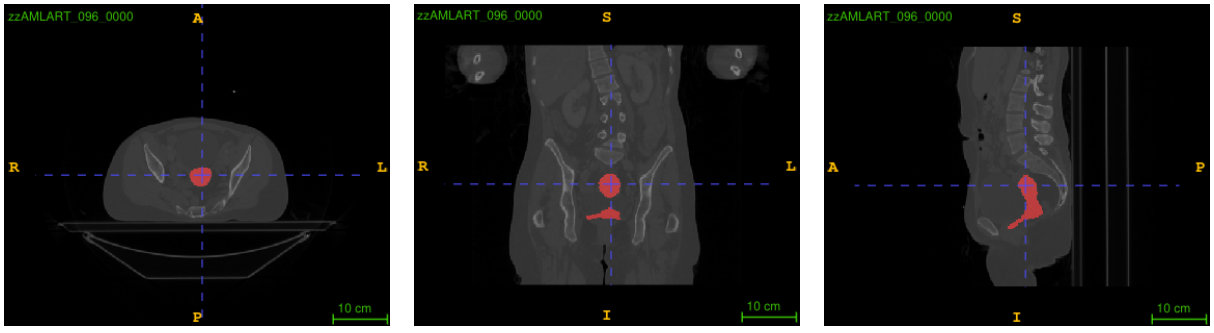
(d) Axial (e) Coronal (f) Sagittal

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

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

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

3.4.2 CTVp



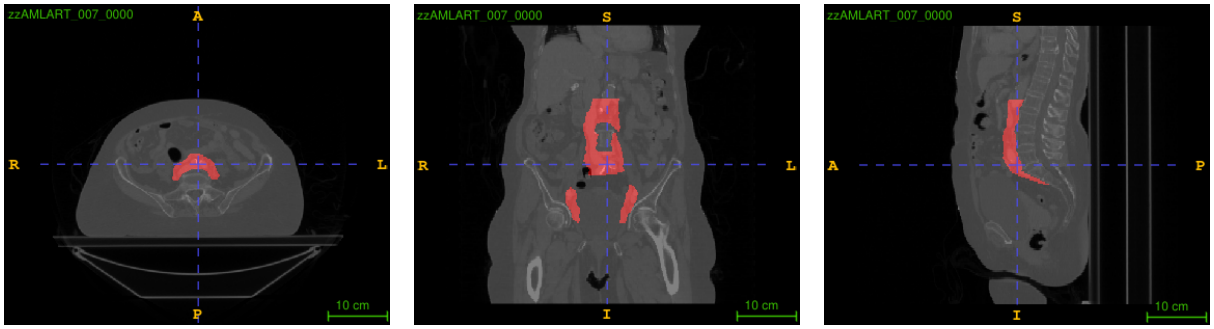
(a) Axial (b) Coronal (c) Sagittal

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

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

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

3.4.3 CTVn



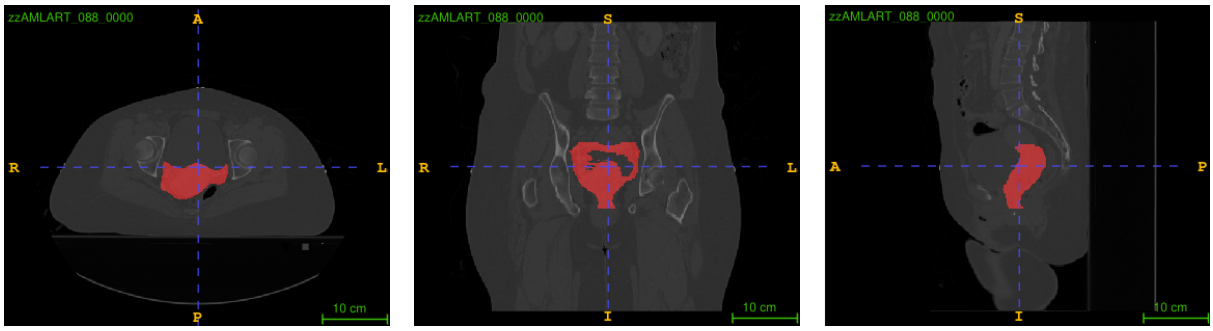
(a) Axial (b) Coronal (c) Sagittal

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

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

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

3.4.4 Parametrium



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

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

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

3.5 Establishing Rules for Structures

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

3.5.1 Relationship between Structures

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

3.5.2 Rules

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

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

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

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

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

3. The Bladder may overlap with the CTVn.

$$B \cap C_n \neq \emptyset \vee B \cap C_n = \emptyset \tag{3.3}$$

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

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

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

$$C_n = G_n \cup L_i \cup L_p + L_{pa} \tag{3.5}$$

Chapter 4

Evaluation Metrics

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

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

However, it is far harder to decide on a scoring system when in a delineation context. Consider a single slice of a CT-scan with a known contour around the perimeter of a tumour[1]. A model like those mentioned in Section 2.2 would attempt to learn a function to closely replicate the contour. Here our domain, prediction and ground truth are all 2-dimensional objects.

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

4.1 Classification Based

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

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

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

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

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

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

4.2 Spatial Overlap Based

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

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

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

A specific case of this equation with $\beta = 1$ is mathematically equivalent to the DICE Similarity Coefficient which was found to be the most popular evaluation metric amongst 2021 studies [**review-metrics**, **evaluation-metrics**, **Sherer2021-le**].

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

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

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

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

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

4.3 Surface Based

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

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

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

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

4.4 Volume Based

Volume-based metrics consider only the volume of the segmentation [**evaluation-of-metrics-in-prostate**, **review-metrics**, **boundary-overlap-metrics**]. However, due to its poor spatial descriptions it is more commonly used jointly with other metrics.

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

4.5 Evaluation

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

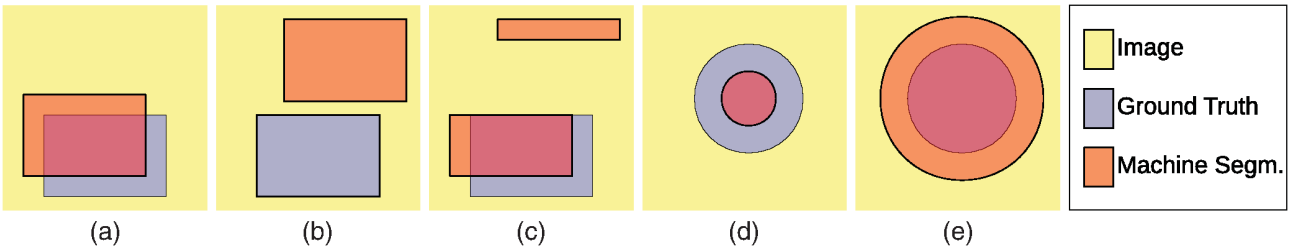


Figure 4.1: Figure from [**boundary-overlap-metrics**] illustrating cases of segmentation to aid with explanation of set-backs of certain evaluation metrics

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

4.6 Estimated Editing Based

> ☞ This is a quotation from this paper, [**Sherer2021-le**], however, it is referencing a paper of its own. Shall I reference the original paper or are 'linked' references OK?

It is difficult to select a measurement which can reflect a clinicians acceptability score. A study found that there was a lack of correlation between a geometric index and expert evaluation, with the JAC score having a 13% False Positive Rate. The conclusion of the study summarised that scores such as JSC and volumetric DSC, "provide limited clinical context and correlation with clinical or dosimetric quality" [**Sherer2021-le**].

Because of the clinical context of evaluating the segmentation by a machine, it may sometimes be helpful to define a performance metric as the "fraction of the surface that needs to be redrawn" [**Nikolov2021-xe**] since models at this point require manual review to avoid automation bias (Section 5.2). For larger structures, this method is useful it doesn't assign a lot of weight on the large trivial internal volume which accounts for a much larger proportion of the score.

4.6.1 Surface DSC

The study at [**Sherer2021-le**] helped drive an initiative to combine aspects of Surface Based evaluation (Section 4.3) and Spatial Overlap Based evaluation (Section 4.2) into a Surface DICE. This

assesses the specified tolerance instead of the overlap of the two volumes.

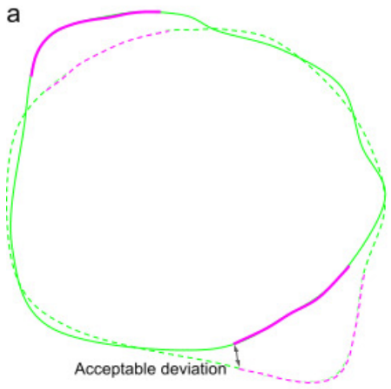


Figure 4.2: Taken from [**Nikolov2021-xe**]. Illustrates the computation of the surface DICE, where the continuous line is the predicted surface and the dashed line is the ground truth. The black arrows show the maximum deviation tolerated without penalty; therefore, in pink is the unacceptable deviations and green otherwise.

We can formulate the Surface DSC score in a mathematical definition [**Sherer2021-le**].

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

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

4.6.2 Added Path Length

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

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

4.7 Summary

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

For the purpose of this project, we shall select a evaluation measurement which is more bias towards conservative boundary estimates to not touch the organs at risk. This choice was in-part influenced by the clinician's review pipeline; it would easier to correct Figure 4.1(d) instead of (e) because correcting the latter would likely take a considerable amount of time as it would require redrawing almost all of the boundary, whereas the former could be corrected much faster [**Nikolov2021-xe**].

Chapter 5

Ethics

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

5.1 Patient disclosures

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

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

5.2 Using the tool

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

Yet, it is too early to take result the medical tool as gospel. For current cervical radiotherapy delineation tools, only 90% of the output is considered as acceptable for clinical use [**auto-delineation-cervical-cancer-**
The remainder therefore has the potential to cause more harm than good if not checked properly. For example, overlap of a PTV onto an organ-at-risk may invoke a cascade of negative effects for the patient. A potential cause may be the lack of multivariate analysis, where an oncologist would need to consider a variety of data, whereas this model only considers a single point of evidence (results of an imaging modality).

Clinicians can fall into the trap of automation-bias as AI becomes more common place in clinical envi-

ronments [**automation-bias**]. However, many models of this age codify the existing bias in common cases, which often will fail those patients who do not fit the expectations of the majority. Therefore, a degree of supervision required from physicians has to be established if this tool is to be used in practice. Oncologists will be required to reverse-engineer results of the 'black-box' to verify why a decision has been made. Secondly, the responsible party for incorrect decisions made by DL tools should also be determined [**AI-in-cancer-diagnosis-era**].

Chapter 6

Interim Deliverables

6.1 Project Plan

The current state of the project has provided me with a very strong foundation for the next 5 months. I have completed detailed research on large foundational concepts with respect to Transfer Learning (TL) principles. I am anxious to include more research however, because TL is only one solution, as a fall-back device it would be beneficial to discuss other techniques as well as provide some quantitative evidence to show shortcomings or strengths.

Figure 6.1, 6.2(a), 6.2(b) is a Gantt chart of the planned progress of the key milestones of the project for the future. I have split tasks into 3 categories: Source code for the programming of the project, Report for the write-up, and the presentation at the end of the project life expectancy.

I expect to consistently document my progress in my final report so that the repository stays most up-to-date with the progress. Furthermore, because of a busy term, I must we aware of slower progress. I will continue pushing with the Transfer Learning approach, and plan to get the bulk work done after the exam period to consider alternative solutions such as 'shall I delineate classes independently' or 'are there better data optimizations or image segmentation approaches'?

Furthermore, I will be attending the weekly meetings to keep the supervisor (Ben Glocker) informed on the progress and receive the most up-to-date guidance on next steps. In addition to supervisor meetings, I must plan an opportunity to meet with staff from the Royal Marsden Hospital in order to discuss clinical details and project expectations as they are the expected primary stake-holders.

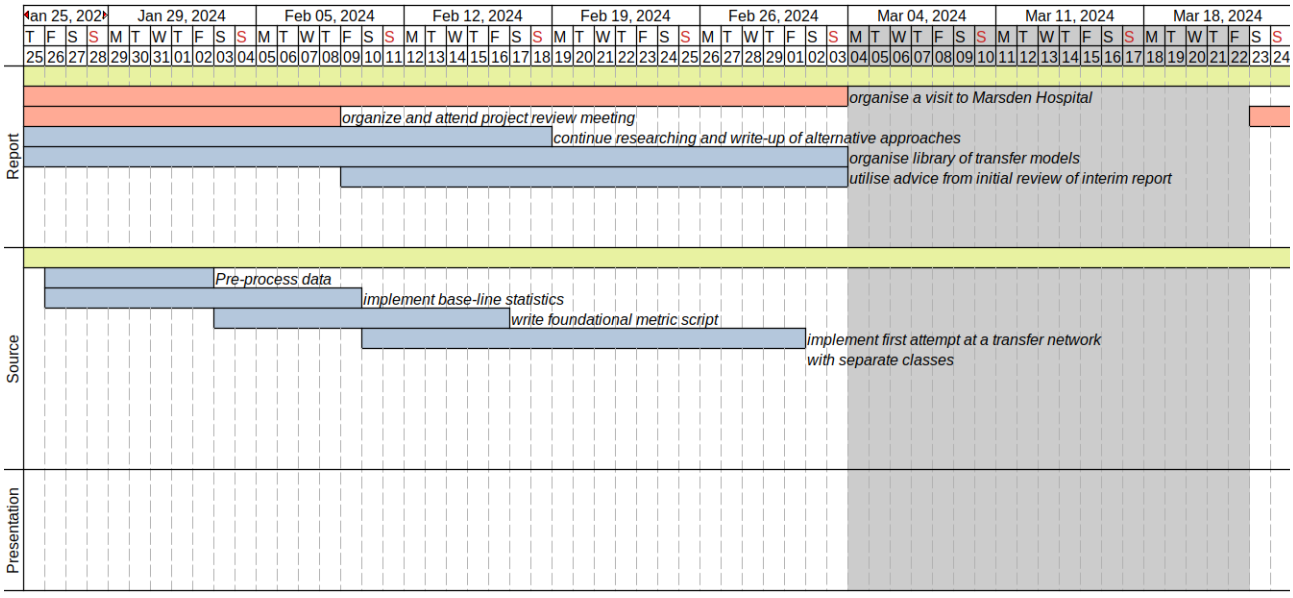
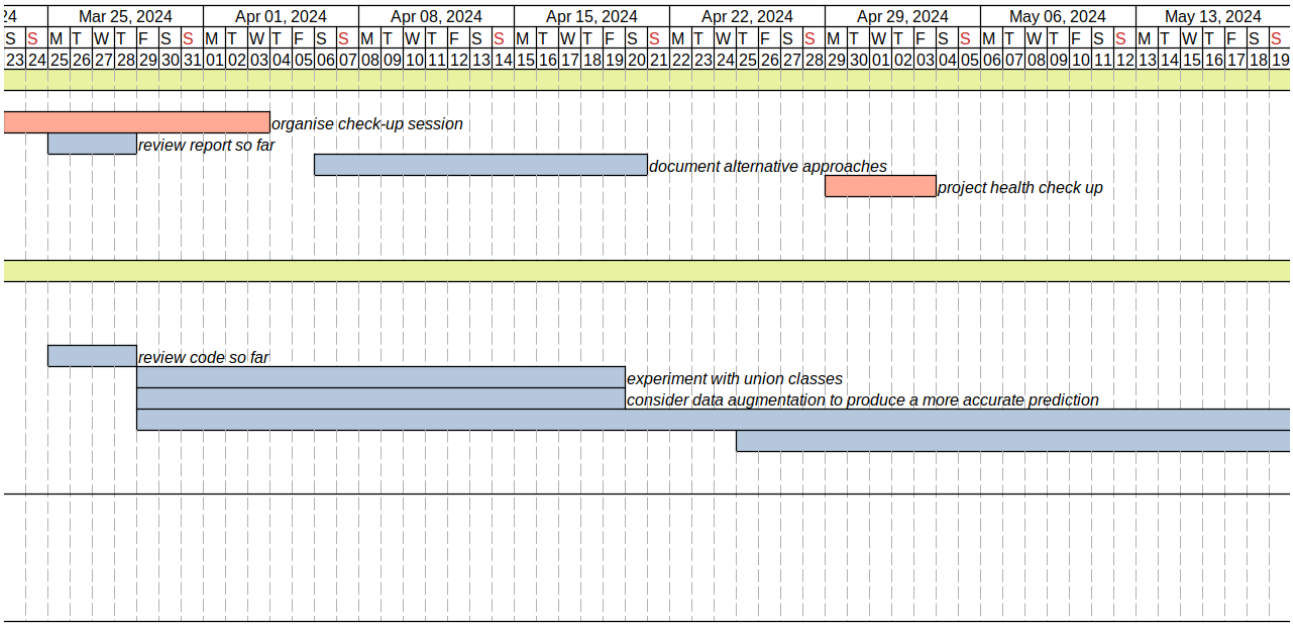
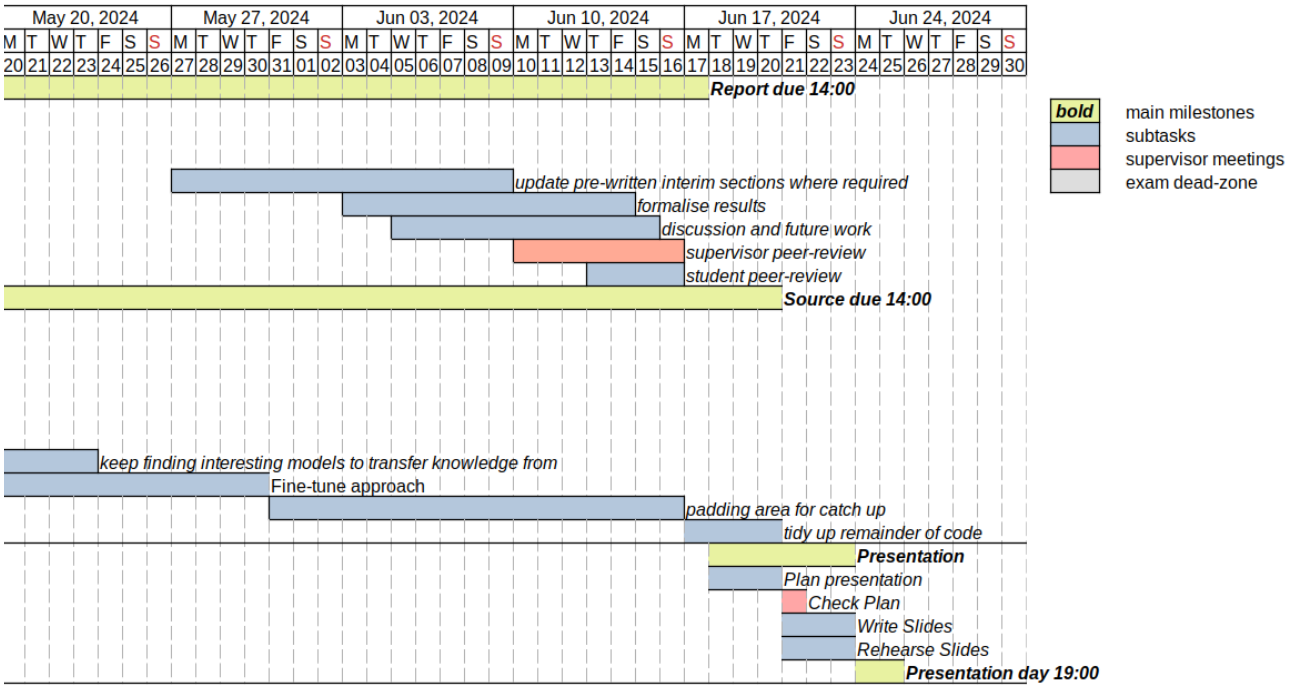


Figure 6.1: Plan for project before exams



(a) Plan for project after exams



(b) Plan for project after exams, the third entry is timeline for the project

Figure 6.1: Gantt chart for the file expectancy of this project

6.2 Evaluation Plan

This project has a convenient way to measure performance metrics of the algorithm; these have been defined in Chapter 4. The project will be considered a success if the proposed solution beats the performance of vanilla network implementations or we can propose reasons for the model's poor performance. Further evaluation will be peer-reviewed by clinicians at the Royal Marsden Hospital, thus reinforcing or alternatively contradicting performance metrics provided.

We will be demonstrating high scoring segmentations according to the provided labelled data from the Royal Marsden Hospital. Current project direction seems to be aimed at reducing the extremities of

editing contour lines form the model to predictions that will be used in clinical contexts. These models will be produced as an consequence of several experiments for fine-tuning the network to have better performance, but yet not generalizing on the data too much.

Finally, the model will be offered as an agent to advise clinicians (not replacing) on radiotherapy contour outlines. The tool should be easy to use, as a simple plug-and-play for the intended stakeholders.