THE UNIVERSITY OF
SYDNEY

Clinical application of deep learning: Automatic contouring via U-net architecture

Matthew Cooper

Supervised by:

Simon Biggs, Riverina Cancer Care Centre

Dr. Yu Sun, University of Sydney

Matthew Sobolewski, Riverina Cancer Care Centre

In partial fulfilment of the requirements for the degree Master of Medical Physics

Institute of Medical Physics

School of Physics

The University of Sydney

June 2020

Contents

Abstract

Purpose: Accurate contouring is a critical aspect of safe and effective treatment delivery in radiotherapy (RT). Current limitation in clinical practice include: Large inter and intra-observer variances, as well time delays in both contour generation and correction - that act as barriers to the implementation of adaptive RT. This study designs and evaluates a 2D U-net architecture with two primary aims: 1) Develop a pelvic imaging quality assurance tool for use in RT, comparing model predictions with expert contours. 2) Automate vacuum bag segmentation for canine RT.

Method: We present two independently trained models in this study, and assess the performance of common semantic segmentation loss functions in each case. We expand on the original architecture developed by Ronneberger et al. by integrating recent network modifications that have shown improved performance in the literature. In addition to reporting dice similarity (DSC), we utilise organ-specific tolerances representative of expert IOV for the bladder and rectum as parameters in Nikolov et al's surface dice similarity coefficient (sDSC).

Results: 3 contours were produced from pelvic imaging CT scans: Patient contours were measured with mean DSC 0.99(0.01), bladder contours with mean DSC 0.86(0.22) and mean sDSC 0.87(0.18), and rectum contours with mean DSC 0.67(0.12) and mean sDSC 0.92(0.14). Additionally, vacuum bag contours from canine imaging CT scans were measured with mean DSC 0.95(0.01). Weighted DSC was the only loss function that optimised for all organs considered in pelvic imaging - due to a significant class imbalance.

Conclusion: Patient contours were excellent. We suspect a broader dataset may improve bladder and rectum segmentations (19 patient scans were used). The vacuum bag model should proceed to acceptance testing for clinical implementation.

Acknowledgements

The author wishes to acknowledge the following:

Simon Biggs

For his mentorship and support over the course of the project. Simon played a pivotal role in shaping the author's philosophy of code development. In addition, he facilitated the author's first professional connections with the medical physics community.

Dr. Yu Sun

For his encouragement and guidance throughout the Master of Medical Physics degree. Yu Sun has also been key in expanding the author's professional network - increasing the exposure of this project.

Matthew Sobolewski

Matthew provided valuable feedback and guidance on both the presentation and the research report. The author wishes to thank Matthew for supporting the first student research project at Riverina Cancer Care Centre.

The team at Riverina Cancer Care Centre

In particular: Jacob McAloney, Nick Menzies, and Simon Biggs, for providing accommodation in their own homes over the course of the project. This allowed the author to gain formative clinical experience. Additionally, this project would not have been possible without the data provided by Riverina Cancer Care Centre (RCCC) and the Small Animal Specialist Hospital (SASH). The medical physics team at RCCC also provided valuable feedback on the presentation.

Dr. Annette Haworth & Dr. Sianne Oktaria

For going above and beyond to ensure March-June 2020 research students were able to complete their projects on time and remotely during the COVID-19 pandemic.

Dr. Robert Finnegan

For providing detailed feedback on the presentation.

The team at DeepMind

For making public their surface-distance-based performance metrics.

https://github.com/deepmind/surface-distance

Statement of contribution

This project was suggested by Simon Biggs and was presented with the freedom to explore any solution that the author believed would have clinical application. All research was conducted by the author; although, Simon suggested Nikolov et al. as an initial blueprint. All code used in this project (except for common machine learning and computational packages) was written by the author and has been released under an open-source license. Additionally, the author built a custom PC with the GPU used for computation in this project - NVIDIA RTX 2070 Super. Simon Biggs organised access to clinical data on site at Riverina Cancer Care Centre (RCCC), and provided a software solution for anonymising patient scans. Furthermore, Simon Biggs organised canine imaging data from the Small Animal Specialist Hospital (SASH) after suggesting that this would be a good place to start getting familiar with coding libraries. In addition to the above, Simon Biggs and Yu Sun provided weekly to bi-weekly feedback on the project direction over the course of the semester.

Code repository: github.com/pymedphys

I certify that this report contains work carried out by myself except where otherwise stated.

- Matthew Cooper 10/06/2020

List of Figures

List of Tables

1 Introduction

1.1 Background

Approximately 18 million patients are diagnosed with cancer each year [1], with reports indicating that up to 50% of all cases could benefit from radiotherapy (RT) as curative or palliative management of disease (optimal radiotherapy utilisation rate) [2]. Contouring is a critical aspect of RT treatment planning; it describes the process of defining and classifying anatomical regions-of-interest (ROIs) within a patient from medical imaging data [**CITATION**]. Contoured regions include: target volumes for treatment, associated error margins; as well as normal tissue regions with differing radio-sensitivities (organs at risk - OARs) - for which exposure needs to be minimised to avoid adverse side-effects of treatment [**CITATION**]. Once ROIs are defined, beam output can be translated to patient anatomy, allowing for dose distributions to be determined for each region (as seen in Figure 1.1). Therefore, contours are part of the primary geometry used to optimise treatment outcomes; achieved by maximising tumour control probability and minimising normal tissue complication probability - i.e. maximising the therapeutic ratio of treatment by optimally distributing dose [**CITATION**]. As such, accurate contouring is fundamental to the efficacy of RT [**CITATION**].

Cite

Cite

Cite

Cite

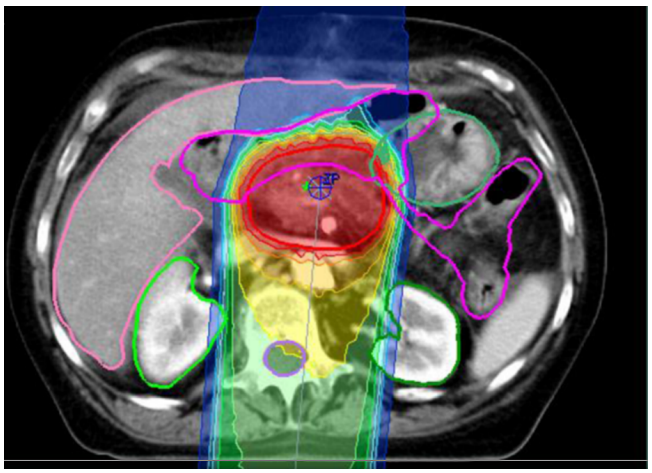


Figure 1.1: Single posterior field setup for carbon ion radiotherapy treatment of pancreatic cancer. Multiple contours are outlined on diagnostic CT imaging for treatment planning. Colour map shows the dose distribution over patient anatomy [3].

However, there are current limitations in clinical practice. Large intra- and inter-practitioner variability (IOV) exists in the definition of ROIs on medical imaging. IOV is a long-standing challenge in RT, and is frequently reported as the largest source of error in accurate treatment delivery [4]. Additionally, manual contouring is both time consuming and requires highly skilled experts for consistency [5]. For instance, research has estimated that a radiation oncologist (RO) needs between 90-120 minutes to delineate pelvic OARs in a cervical cancer patient [6]. The process is often computer aided and automatic OAR segmentation tools (i.e. deformable image registration or Atlas-based methods) have both reduced contouring times and

improved consistency between expert observers [4]. However, atlas-based methods still require significant manual editing [5], and experience difficulty with small organ volumes, regions with poor contrast for differentiation, or high variability in size or location - such as pelvic OARs in prostate and cervical cancer [7, 6]. Critically, current automatic solutions still present a barrier to the adoption of future technologies that would require fast contouring [5]. For instance, adaptive radiotherapy has shown the potential to deliver a new standard of care for RT patients by updating treatment plans to daily changes in patient anatomy [5].

In contrast, deep-learning (DL) algorithms have shown significant performance improvements over atlas-methods both in terms of accuracy and time for contour generation. However, the implementation of DL models in clinical environments remains challenging; particularly due to limitations in quantitatively assessing model performance in comparison to expert performance and IOV [5]. Typical metrics used to quantify the similarity between model and expert contours are volumetric in nature, hence volume overlap tends to be the focus of evaluation [5]. Recent studies have introduced surface-based performance metrics that aim to provide direct information on the fraction of surface points that require correction to be within IOV tolerances (specific to each OAR) [5, 8], and may provide a stronger correlation with time required for contour correction [8].

1.2 Aim

U-net is a type of deep-learning architecture that leverages multi-resolution analysis to perform state-of-the-art segmentation in medical imaging research [9, 10]. We present two independent U-net models in this study, and expand on the original architecture developed by Ronneberger et. al in [11], by integrating recent network modifications that have shown improved performance in the literature:

Model 1 - Pelvic imaging QA tool

Model 1 was designed to fulfil the need for contouring to become part of regular quality assurance due to large IOVs reported across the literature [4]. In addition, model 1 aims to evaluate the ability of U-net to achieve expert level performance - as defined by Nikolov et. al's surface-based performance metric (surface dice similarity coefficient) [5]. We focused on pelvic CT imaging scans, automatically contouring the patient, bladder and rectum volumes. The model then compares predicted contours with those created by an expert clinician; and provides feedback on the volumetric overlap (dice similarity coefficient - DSC), as well as the percentage of surface points that deviate by a distance larger than the IOV associated with each structure (surface dice similarity coefficient - sDSC). Recent studies have highlighted that surface overlap metrics - such as the sDSC - may have improved correlation with respect to manual correction times required for an automatic segmentation, when compared with standard volumetric overlap metrics [8]. We present both metrics for comparison, with models trained over a variety of loss functions common to medical image segmentation.

Model 2 - Automatic segmentation of vacuum bag structures in canine imaging

Model 2 was designed to automatically contour vacuum bag structure in canine imaging. Vacuum bag structures are reported to take approximately 30 minutes per patient to contour [**CITATION**]; hence, this model aims to automate a time-consuming aspect of RT treatment that is typically processed manually in veterinary medicine.

2 Literature

2.1 Introduction

Hardware developments in the last two decades (2000-2020) have enabled deep-learning algorithms to achieve state-of-the-art image segmentation in the field of computer vision, outperforming humans on many classification tasks [12, 13, 14]. The following review aims to cover recent developments in medical image segmentation by focusing on network architecture, training methods, and challenges in the context of radiotherapy (RT). We begin by examining variability in the definition of organs contoured by medical professionals, and outline a novel technique used by Nikolov et al. that takes this variability into account when assessing model performance [5]. Further, we explain U-net architecture in detail before expanding on the original work by examining modifications in recent implementations. We conclude the review by outlining class imbalance as an optimisation challenge in medical imaging segmentation.

2.2 Observer variability in contour delineation

Uncertainty in the delineation of contour volumes is a significant source of error in radiotherapy (RT) [5]. Multiple studies have highlighted accurate contouring as a requirement for effective clinical outcomes [4, 15, 16], and note that current inter and intra-observer variability creates a challenge for quality assurance (QA) of dosimetric impact [4]. Additionally, inconsistencies in adherence to contouring protocols have the potential to introduce variability when cross-referencing the results of clinical trials [15]. Investigations into the contouring quality of trials have revealed that up to 80% of audited files would require modification for protocol compliance [17]. A separate study found a 25% non-compliance rate in a phase 3 trial for head and neck RT - primarily due to incorrect target contouring - which was associated with a 20% decrease in 2-year survival rates [18]. High-quality contours are often achieved through a combination of highly skilled multidisciplinary teams, taking into account additional data beyond patient imaging [4, 15]. However, in attempts to decrease variability and its impact on treatment outcomes, studies report the benefit of automatic contouring tools as a starting reference (when available) [4], and highlight the need for volume delineation to become part of routine QA [4].

While automatic contouring has been shown to improve consistency in delineation [4], current commercial solutions do not provide a fully automated experience [16]. For instance, Varian Medical Systems 'Smart Segmentation' tool includes the trachea and main bronchi in normal lung delineations, contrary to the RTOG 1106 guideline for lung cancer RT [16]. Hence, there is an urgent need for alternative contouring tools in RT that provide more accurate delineation [16], and feedback on the accuracy of manual corrections for QA [5].

In RT, inter-observer variation (IOV) is often larger than errors associated with patient setup and organ motion [4, 19]; however, the extent is also organ dependent [15]. A 2019 study measured IOV on diagnostic computed tomography (CT) for prostate cancer treatment and found "excellent agreement" for bladder, rectum, and clinical target volume contours (defined by an inter-observer variability assessment (ICC) value > 0.75), and reported the extent of

variation to serve as a benchmark for comparison [15]. A total of 5 patients were included in the study, selected to represent the broad range of anatomy within clinical trials. Contouring was performed by 13 observers (9 radiation oncologists) across multiple clinic locations, with guideline examples distributed to all observers before study commencement. Bladder volume agreement was measured to be 0.93 ± 0.03 via the dice similarity coefficient (DSC [20] - see equation 2.2 and Figure 2.1), with absolute mean surface distance (MSD) ranging from 0.76 mm to 1.44 mm across patients [15]. DSC agreement for the rectum was 0.81 ± 0.07, and a MSD of 1.97mm to 4.14mm [15]. Although variance for the rectum was higher than that of the bladder, external studies report that a DSC ≥ 0.7 is considered clinically acceptable for these organs [15, 21]. Consistent IOV is reported across the literature for male bladder and rectum contouring via diagnostic CT [22].

2.3 Defining an expert performance metric for model evaluation

Although automated segmentation tools are in clinical use [10], current performance tends to be poor when compared to expert delineation - especially in the case of small organ segmentation [5, 10]. Consequentially, they are typically used as a starting reference and require time-consuming corrections [5, 16]. A study developed by Google's DeepMind highlights an additional barrier to deep-learning solutions (which have shown significant improvements in accuracy and inference time over traditional methods [10]) is the absence of a clinically-relevant metric that takes into account expert IOV when assessing model performance [5]. The study suggests that the typical DSC metric is a poor measure of similarity between model and expert if operating under the assumption that manual correction is required [5]. DSC is a commonly used volumetric overlap score (equation 2.2) in medical image segmentation [23]. However, DSC does not penalise a model for the number of contour surface points that must be manually adjusted [5] - which may be an important indicator of time required for correction [5]. In an attempt to overcome these limitations, Nikolov et al. introduced a novel 'surface DSC' metric (sDSC), which measures the similarity between two contour boundaries and normalises to be independent of organ volume. This metric defines expert performance τ as the 95th percentile MSD between observers (specific to each organ considered) and enforces no penalty when surface deviations are within this tolerance [5]. In contrast to the DSC metric, surface DSC measures the proportion of surfaces in a contour set which are within expert IOV; providing direct information on the degree of manual correction required [5]. Successful model performance was defined by Nikolov et al. as a cross-patient average sDSC $\geq 95\%$. An illustration of the proposed metric is included in Figure 2.1, where M_i represents the volumetric mask considered in DSC measures, $B_i^{(\tau)}$ represents the contour surface S_i with IOV tolerance τ, and $S_i \cap B_j^{(\tau)}$ is the intersection of surface boundaries at an organ-specific tolerance τ. The sDSC metric is stated in equation 2.1 [16], while the standard DSC is included for comparison in equation 2.2 [24].

$$sDSC_{i,j}^{(\tau)} = \frac{|S_i \cap B_j^{(\tau)}| + |S_j \cap B_i^{(\tau)}|}{|S_i| + |S_j|} \tag{2.1}$$

$$DSC_{i,j} = \frac{2|M_i \cap M_j|}{|M_i| + |M_j|} \tag{2.2}$$

However, the proposed metric has yet to see clinical implementation. As such, Nikolov et al. recommend reporting sDSC values alongside typical DSC measures to compare performance across the literature [5]. An additional limitation in the use of sDSC (a so-called 'hard' metric) is the introduction of a disharmony between model optimisation and target performance [24].

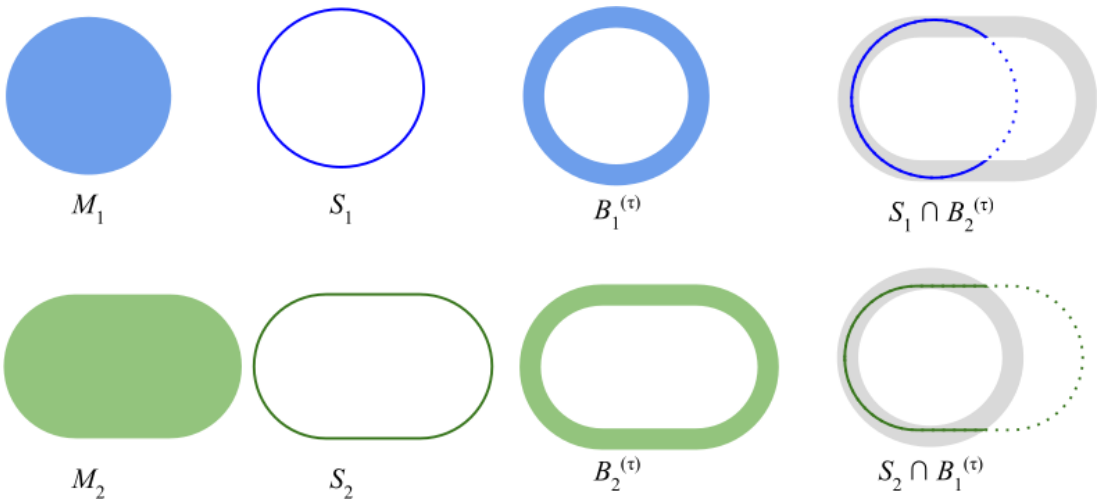


Figure 2.1: Illustration of equation variables seen in DeepMind's proposed surface dice similarity coefficient (sDSC), equation 2.1; and typical dice similarity coefficient (DSC), equation 2.2 for contours i, j. M_i represents the volumetric mask considered in DSC measures, $B_i^{(\tau)}$ represents the contour surface S_i with inter-observer variance (IOV) tolerance τ, and $S_i \cap B_j^{(\tau)}$ is the intersection of surface boundaries at organ specific tolerance τ: defined as the 95^{th} percentile absolute mean surface distance (MSD) between expert observer contours, specific to each organ considered [5].

Hard metrics assume binary segmentation as input; therefore, soft surrogates (accepting continuous data) are required to define differentials for gradient descent algorithms [24]. Recent studies have proposed soft DSC as a loss function that can directly optimise the DSC performance metric [24]; however, no such surrogate currently exists for the sDSC due to difficulties in defining surface integrals on boundaries represented by continuous segmentation values. Multiple sources in the literature have highlighted the benefit of directly optimising for the metric used to evaluate model performance [24, 25].

Vaassen et. al attempted to measure the relationship between the sDSC and time required for contour correction [26], as seen in Figure 2.2. However, this study failed to use a 95th percentile MSD for each organ-specific tolerance, as described by Nikolov et. al. Rather, a value of 1 mm was used for all organs at risk (OARs), corresponding to the x-y pixel resolution used in the study's CT scans [26]. Hence, sDSC values presented in this study are representative of variance due to spatial resolution limits and not IOV. MSD_{95} values calculated from Roach et al. show organ-specific tolerances of 1.46 mm for the bladder, and 6.99 mm for the rectum. Thus, we expect sDSC values reported by Vaassen et al. to be significantly lower [5], and have poorer predictive validity than when considered under correct IOV assumptions. Despite this limitation, the sDSC was a significantly better indicator of time required for correction when compared to the DSC and mean Hausdorff distance (MSHD) metrics [26]. To the best of our knowledge, no other attempt to correlate sDSC with correction time exists within literature.

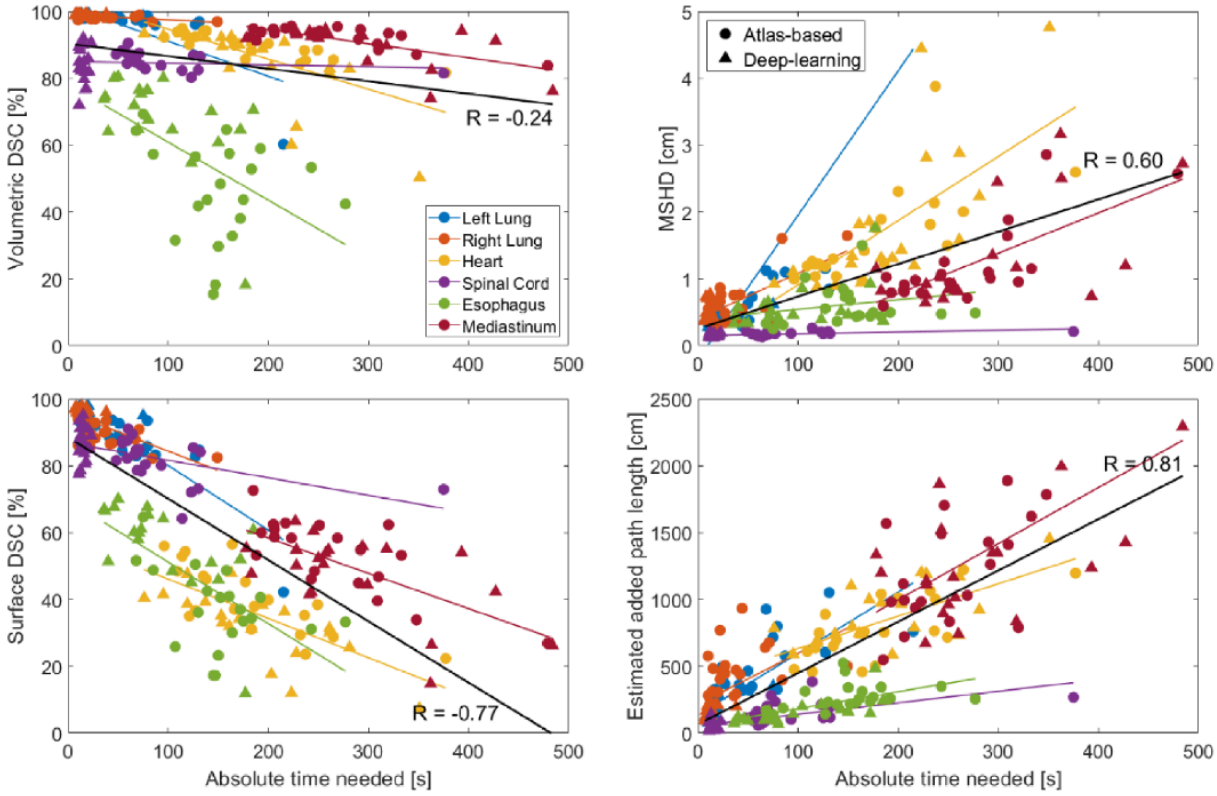


Figure 2.2: Vaassen et. al compare common segmentation similarity metrics with surface DSC (sDSC [5]) and their novel 'estimated added path length' metric for ability to infer absolute time required for automatic contour correction. Atlas-based (circles) and deep-learning (triangles) methods combined. Correlation coefficients indicate a stronger relationship between sDSC value and time required, than dice similarity coefficient (DSC) and mean Hausdorff distance (MSHD) [8]. We note a limitation to this study is the use of an incorrect organ specific tolerance (1.0 mm - voxel size), compared to the organ specific inter-observer variance tolerance τ defined in Nikolov et. al [5].

2.4 Historical U-net architecture

A 2015 study by Ronneberger et al. expanded on the concept of fully convolutional networks (FCNs by Long et al. [27]) to meet the challenges posed by a lack of curated training data for biomedical image segmentation [11]; and to tailor FCNs for pixel-wise classification (semantic segmentation) [28]. This breakthrough 'U-net' architecture (Figure 2.3) includes a contracting pathway (typical of past FCNs) which downsamples resolution in the image plane to capture contextual (high-level) features for detection and general localisation [16]; as well as an expanding pathway, whereby high-resolution feature maps are concatenated with matching upsampling blocks via residual skip connections [29] to recover full spatial resolution in the model output [28]. Studies report concatenating high-resolution feature maps via this secondary path improves local (detailed) feature propagation [16], as later convolutions operate over both the high-resolution information provided by skip connections, and the contextual features passed via upsampling. Hence, U-net architecture facilitates multi-resolution analysis [29] in an attempt to overcome the trade-off between local feature propagation and the use of contextual information in segmentation [30]. For instance, larger kernel sizes relative to the input resolution infer spatially broader information; although, require additional pooling layers, decreasing local accuracy for detailed segmentation borders [30]. However, studies have indicated that this trade-off still exists, as shallow U-net models tend to perform better on small segmentation regions [10].

Contracting (encoding) blocks were composed of two convolutional layers with 3 x 3 kernel sizing, doubling the feature channels of the input before downsampling the x-y resolution via 2 x 2 max-pooling [11]. External studies have demonstrated the importance of increasing channels before max-pooling operations to avoid computational bottlenecks [31]; a strategy adopted in both paths of typical U-net models [31]. In contrast, each expanding (decoding) block upsamples resolution before halving the number of feature channels by successive 3 x 3 convolution [11]. As this architecture does not make use of convolutional padding, skip connections must be cropped [11], resulting in a segmentation mask with reduced size when compared to the input image. State-of-the-art networks make use of mirrored padding to overcome this limitation [5].

2.4.1 Activation functions

Activation functions present in U-net introduce non-linearity into the model, to enable learning of sophisticated features, beyond those extractable by matrix multiplication alone [29]. Additionally, ReLU decreases the computational burden compared to typical activation functions (such as tanh and sigmoid), whilst preserving the requirement for non-linearity [32]. However, activation design is an ongoing area of research, with some studies noting performance improvements under modified ReLU functions (i.e. LeakyReLU) [33]. Additional theory on activation functions is provided in section 3.2.5.

2.4.2 Data augmentation

Pre-processing for U-net architecture routinely incorporates data augmentation, to add biologically relevant sources of variation to the training data [29, 30, 34]. Thus, improving both model robustness and data-use efficiency [11]. Dosovitskiy et al. demonstrated improved inference reliability after increasing their effective dataset size via geometric transformations, voxel intensity modulation, as well as blur and noise filters [35]. Data augmentation is also applied to balance the number of infrequent labels in a biased dataset [29] - such as those in medical imaging, where relatively smaller ROIs occupy a limited volume of the total contour space [36].

2.5 Current U-net architectures

U-net implementations and applications vary broadly across the literature - from simplified 2D versions with 3 downsampling layers and 4 x 4 convolutions [16], to 3D models that accept full patient volumes as input [10]. A theoretical benefit of volumetric input is that organs may have axial markers that would be absent from 2D images [30]; model inference may benefit from the richer spatial information provided by 3D inputs [16]. For example, Nikolov et al. presented a 3D U-net architecture for head and neck segmentation that accepted multiple context slices in addition to the primary input image [5] - delivering clinically acceptable contours over most datasets, for all but the smallest organs considered [5]. Alternative 3D implementations report promising results [5, 10, 37]. However, studies designed to quantify improvement over the simpler 2D architecture have shown limited performance gains [16]. In contrast, 3D networks require a significantly higher degree of computational resources, which may pose a challenge for clinical implementation [16]. In general, state-of-the-art 3D networks are trained and deployed via cloud services [16]; raising important ethical questions for clinics, where data

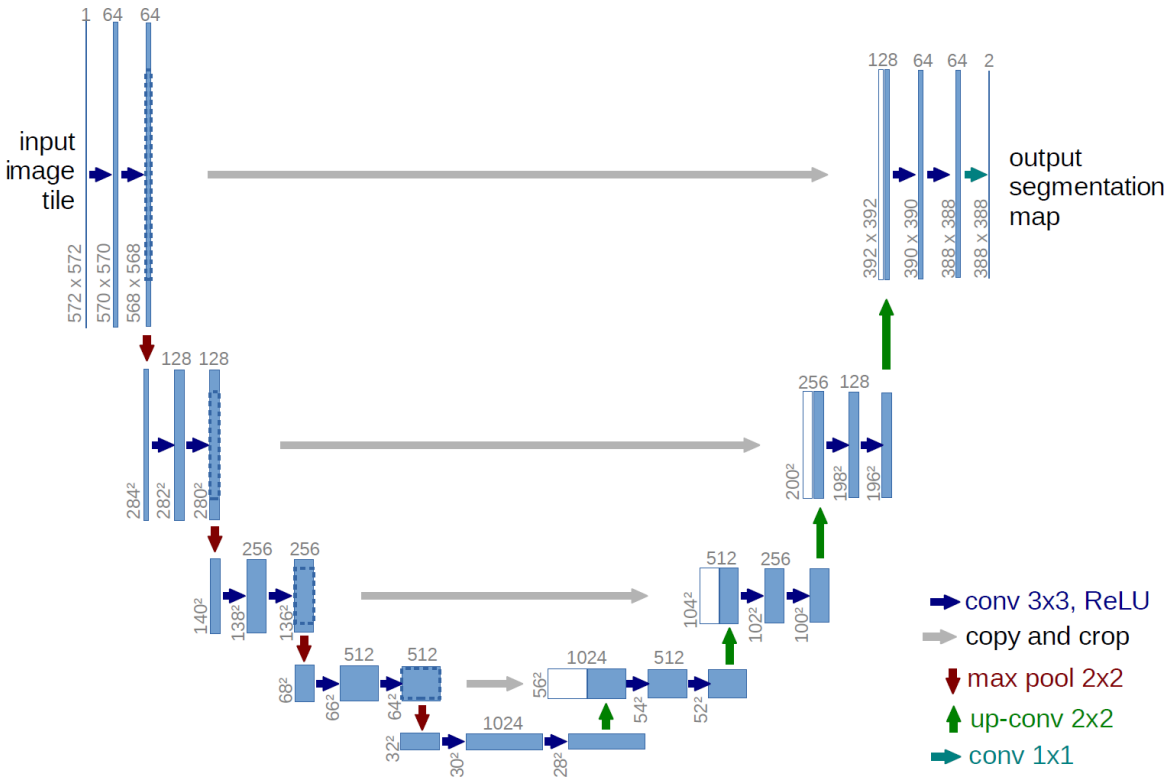


Figure 2.3: Original U-net architecture first proposed in 2015 by Ronneberger et al. [11]. The model consists of symmetric encoding (left) and decoding (right) pathways. Residual skip connections allow for concatenation of extracted image features at different resolutions in order to provide both high-level localisation and high resolution local information for accurate segmentation [16].

is typically stored on-site with strict security and privacy protocols in place [16, 34]. In addition, modified networks have experimented with feature summation between up-sampling and skip connections; however, concatenation (as in Ronneberger et al [11]) has shown consistently better performance [10].

Nemoto et al. showed that both 2D and 3D U-net implementations were more effective than commercially available atlas-based auto segmentation tools for delineation of lung regions [16]. A total of 232 patients were selected for model training and testing, with all segmentations determined manually by expert observers. In contrast to previous U-net models mentioned in this review, Nemoto et al. made use of batch normalisation layers (which aim to reduce the effect of bias output distributions from the previous layer to accelerate training [38]). Wilcoxon signed-rank testing showed that performance gains over atlas segmentation for both U-net architectures were statistically significant with $P_{val} < 0.01$, and mean DSC improvement of 2.7% [16]. However, no statistically significant difference was observed between 2D and 3D U-net models [16], indicating that in the case of lung segmentation (a relatively large structure compared with the model output), the addition of axial input data did not translate to improved contours [16].

2.6 State-of-the-art models for bladder and rectum contouring

Deep-learning segmentation has consistently shown significant improvements over traditional techniques: pixel intensity thresholding, and atlas-based image registration [39]. In the case of pelvic imaging, poor contrast due to similar CT numbers between OARs and high variation in both location and size across patient cohorts restrict the utility of intensity and atlas-based

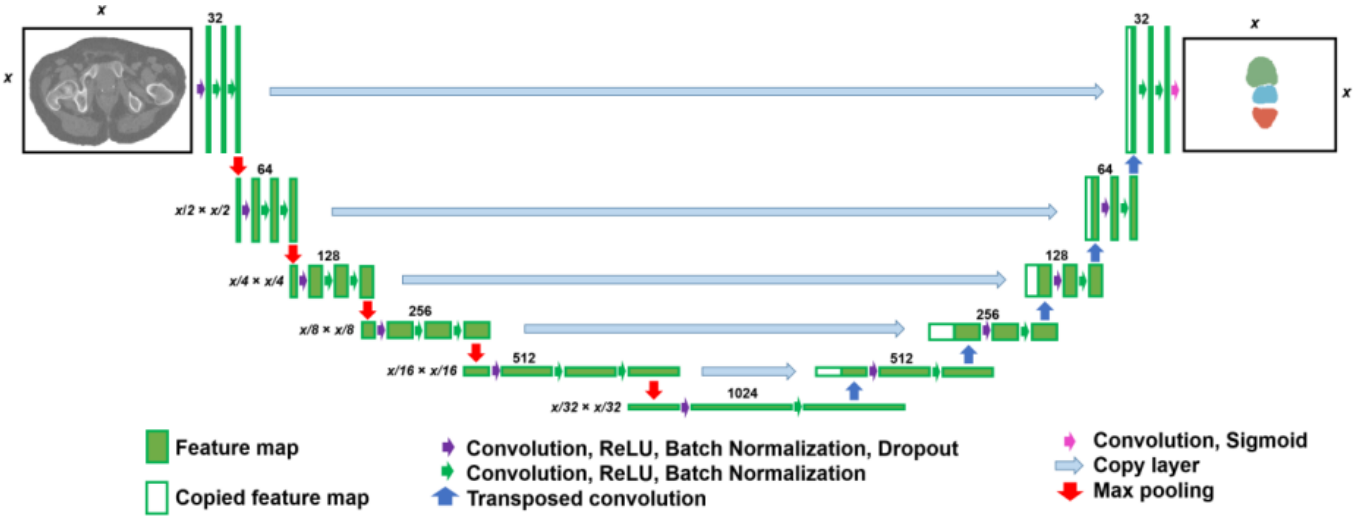


Figure 2.4: Modified U-net architecture used by Kazemifar et. al for state-of-the-art bladder and rectum segmentation in pelvic imaging [9]. Addition of batch normalisation layers, increasing dropout rates, and additional downsampling block; when compared to the original U-net model by Ronneberger et. al. Sigmoid activation used as final layer in the network, compared with in-loss function calculation as used in Ronneberger et. al [11].

methods [40]. In addition to this, atlas approaches tend to have reduced performance in the segmentation of small volume organs [40]. Ayadi et al. conducted a multi-centre prostate cancer study to evaluate the performance of a multi-atlas strategy. Expert contours were compared with Elekta's Atlas-Based Auto-segmentation (ABAS) for 26 patients, resulting in average DSC values of 0.80 ± 0.19 for the bladder, and 0.66 ± 0.09 for the rectum - requiring significant manual correction for use in RT [41].

In comparison, Liu et al. examined the performance of the original U-net architecture in the segmentation of pelvic OARs. Average DSC scores for CT segmentation of 105 patients were reported as 0.90 ± 0.11 and 0.78 ± 0.03 for the bladder and rectum, respectively [6]. This is comparatively lower than 2D U-net results presented by Kazemifar et al. (see Figure 2.4 for schematic) who modified the original architecture via batch normalisation, increasing dropout rates, stochastic gradient descent optimisation (SGD) rather than adaptive moment optimisation (Adam), and an additional downsampling block [9]. Using a DSC surrogate loss function, excellent agreement was found between observer and model contours for the 85 prostate cancer patients included in the study; with average DSCs of 0.95 ± 0.04 and 0.92 ± 0.06 for the bladder and rectum [9]. In a follow-up study, Bologapal and Kazemifar et al. attempted 3D U-net architecture with an additional ResNet block (designed to improve optimisation on extremely deep networks, by minimising numerically unstable gradient flow [29] i.e. the degradation problem [42]), in an attempt to further improve bladder and rectum segmentation. Results showed a larger deviation between expert and model contours for the rectum, with a DSC of 0.95 ± 1.5 and 0.84 ± 3.7 for bladder and prostate, respectively [43].

Finally, Wong et al. evaluated a commercial deep-learning package (Limbus Contour) based on U-net architecture and created an independent model for each OAR to be segmented, trained on an average of 328 CT patient scans per model [44]. Specific details of the architecture are unknown due to the closed nature of the codebase [44]; however, dropout and batch normalisation were selected during training [44]. Data augmentation was also used, including image flipping, intensity modulation, and elastic deformations [44]. Total contouring times for bladder, rectum, femoral heads, prostate, and seminal vesicles were recorded for the deep-learning model and compared to times for a single radiation oncologist (RO). The deep-learning model

(DL) showed a significant time improvement (98% reduction) over manual contouring, with an average inference of 0.4 minutes per patient (excluding manual correction), compared to 21.3 minutes for expert contour (EC) [44]. In addition, DL accuracy in pelvic OAR contouring was comparable to average IOV measurements between 3 ROs [44], highlighting the clinical potential of DL contouring methods in RT. Average worst DSC (lowest DSC score for each patient scan) between DL and EC for the bladder was 0.97, with single worst case 0.95. Average worst EC-EC DSC was measured at 0.96, with single worst 0.94 [44]. Rectum values showed similar consistency between DL and EC contours, with a DL-EC average worst DSC of 0.78, and single worst of 0.49. EC-EC average worst DSC of 0.79, and single worst of 0.55 [44]. However, Wong et al. note the study's limitation due to a small patient testing set (20 patients per disease site) [44] - small datasets are a common challenge in the application of DL methods to segmentation tasks in RT [11, 29, 30, 34].

2.7 Class imbalance in medical imaging

A well-known challenge in applying deep-learning methods to medical image segmentation is the class imbalance between regions-of-interest [30]. In particular, anatomical data in CT imaging often integrates to a much smaller volume than the background voxels, and organs inside this volume vary significantly in size [45]. As a consequence of this input imbalance, the majority of voxels in a segmentation mask may be negatives, resulting in a a biased optimisation strategy that favours negative predictions [45]. Multi-organ segmentation faces the additional challenge of an input class imbalance between organs, leading to parameter updates that are dominated by ROIs with the most voxels in the dataset, as these contribute most significantly to the loss calculation [36]. Without accounting for this imbalance, the model can quickly become trapped in a local minimum, where optimisation occurs only for the dominant class [36].

Output imbalance refers to a disparity between false-positive and false negative pixels in model predictions. Depending on the context, reducing false-positives may be more important than false-negatives, or vice-versa [45]. For example by placing a higher penalty on false-positives, a model could be tailored to classify less normal tissue as a treatment volume. Conversely, penalising false-negatives may focus optimisation on difficult OARs with poor boundary contrast, reducing under-segmentation [45].

Studies indicate that the use of class weighting in loss functions can improve results by placing higher importance on contours that infrequently occur in the data, or occupy small anatomical volumes in the patient [45]. Christ et al. showed performance gains by implementing a pixel-frequency class weighting to binary cross-entropy loss, focusing optimisation on difficult regions-of-interest [46]; and observed that accurate small lesion segmentation was not possible without class balancing, as their total contribution was less than 1% of contour voxels [46]. In contrast, Taghanaki et al. used weighting and a combination DSC + BCE loss to enforce a desired trade-off between either false-positives or false-negatives (depending on model requirements) to correct for output imbalance; reporting improved DSC scores and lower false-negative rates for multi-organ segmentation across a range of imaging modalities [45]. Further details of loss strategies to handle imbalance are provide in section 4.3.

3 Theory

3.1 Rule 0 - No magic!

Machine learning (ML) is an iterative process of improving a map between an input variable and an output target [29]. In the case of medical image segmentation, initial model input corresponds to diagnostic patient images, and output to the ground truth segmentations created by an expert clinician [9]. More generally, the map from input to output can be interpreted as a series of pattern recognition tasks (functions) used to automate decision making [29]. 'Training' a model involves repeated exposure to a large representative subset of all available data, and application of an optimisation algorithm designed to extract and select features that have predictive validity [29]. For instance, if a vector $x \in \mathbb{R}^n$ is a complete representation of the features extracted by a model with parameters θ learned during the training phase, $\hat{y} = \hat{f}(\theta; x)$ is the prediction produced under the model $\hat{f}$.[1] The primary decision-making unit of a neural network is the perceptron [29]. Each perceptron includes a set of parameters $\theta = (w_0, ..., w_n)$, where w_0 is referred to as the activation bias, and $w = (w_1, .., w_n)$ as the weights. Together, these parameters perform a linear transformation $z = w \cdot x + w_0$ of the input x. A single perceptron network has the modelling capacity of a binary classifier when combined with a smooth non-linear bounded monotonic function h [29]; i.e. $\hat{f}(\theta; x) = h(z)$ is the perceptron output or activation value of the node ('neuron'). Activation functions - historically a sigmoid function $\sigma(z)$ - ensure gradients are well defined during the optimisation process, which iteratively improves model performance by comparing predicted output with the target output [29].

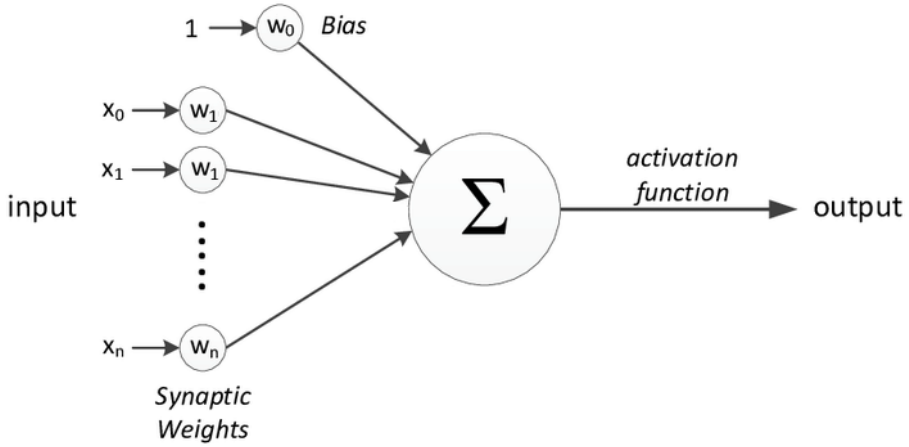


Figure 3.1: Single perceptron example with inputs x, trainable model parameters $\theta = (w_0, ..., w_n)$, and a non-linear activation function h. Output (or neuron activation value) is the 'activated' linear combination $\hat{f}(\theta; x) = h(w \cdot x + w_0)$ [49]. Modified for notional consistency with this document.

[1] Notation: $f(u; v)$ represents a function f with variables u evaluated at a fixed point v. For instance, in the initial input layer v is the model input (CT image) upon which operators with parameters u extract a feature map. This notation highlights that feature maps (also represented by u when used as input in deeper layers) are dependent on the trainable model parameters v.

From here, we already have the fundamental structure required to tell a student if they would pass or fail a subject without prior knowledge of assessment weighting (i.e. a binary classifier for pass/fail) [50]. For instance, by exposing the model to enough past student assessment results x and their final subject score $f(x)$, parameters of the perceptron θ can converge to an approximation of the true assessment weightings [34]. Parameter convergence is achieved via a loss function which calculates the difference between target outputs $f(x)$ under the true map f, and the model predictions $\hat{f}(\theta; x)$ inferred from a 'forward-pass' of the learned approximation $\hat{f}$ [24]. An optimisation algorithm minimises the loss $L(\theta) \sim |f(x) - \hat{f}(\theta; x)| \rightarrow 0$ with respect to θ, while ensuring changes generalise to independent cases (i.e. new student data) from a validation dataset, sampled from the same distribution as the training data (i.e. same subject) [29]. We note here that $L(\theta)$ is a user defined function to measure the difference between prediction and ground truth; and that the geometric norm was presented above for simplicity. Adjustment vectors (gradients) are calculated via the familiar differential operations of calculus (as seen in equation 3.1) [29].

$$\partial_\theta L = \partial_{\hat{f}} L \, \partial_\theta \hat{f} \tag{3.1}$$

Iterative parameter updates converge $L(\theta)$ to a local minimum via a first-order approximation (equation 3.2), commonly referred to as the gradient descent algorithm, where $\alpha > 0$ is the step-size or 'learning-rate' [29].

$$\theta^{i+1} = \theta^i - \alpha \partial_\theta L \tag{3.2}$$

However, it is only through the combination of multiple perceptrons that we can model non-linear decision boundaries [29]. Literature has shown that a single layer multi-perception network is equivalent to an XOR operator [51], and hence can approximate any continuous function on a closed and bounded Euclidean subset (i.e. compact subspace) of $\mathbb{R}^n$ [52]. In multi-layer (indexed from 0 to n) multi-perceptron arrangements (MLPs) i.e. $\hat{f}(\theta; x) = \hat{f}_n(\theta_n; (\ldots \hat{f}_0(\theta_0; x)))$, each neuron accepts activations from the previous layer as input - facilitating the emergence of complex decision-making processes and increased modelling capacity [29]. In this case, the differential vector for loss (equation 3.1) must account for influence through multiple network pathways via the successive application of the chain rule (i.e. the back-propagation algorithm) [29]. An additional complication arises from the fact that updates are usually averaged over a subset (mini-batch) of all available training data to reduce the computational burden (via stochastic gradient descent) [53]; therefore, gradient calculations occur on an approximation of the true loss topology [53]. Still, all we need is calculus, linear algebra, and some additional accounting - no magic!

3.2 Going deeper with convolutional neural networks

Convolutional neural networks (CNNs) are a subcategory of MLP networks, commonly used for deep learning concerning imaging tasks [29], where we can exploit the spatial relationship of input data to reduce complexity when compared to fully connected MLPs [34]. Typical CNN operations consist of convolutional layers that perform feature extraction [30]; as well as pooling layers responsible for feature selection via the sub-sampling of feature maps [11]. Additionally, regularisation constraints (such as dropout layers) prevent over-fitting to the training data - a state by which features extracted under the model parameters (i.e. kernels in the case of CNNs) improve performance on the training data, yet fail to generalise to independent datasets [34]. We provide a survey of each common layer type below, with visual context provided in Figure 3.2.

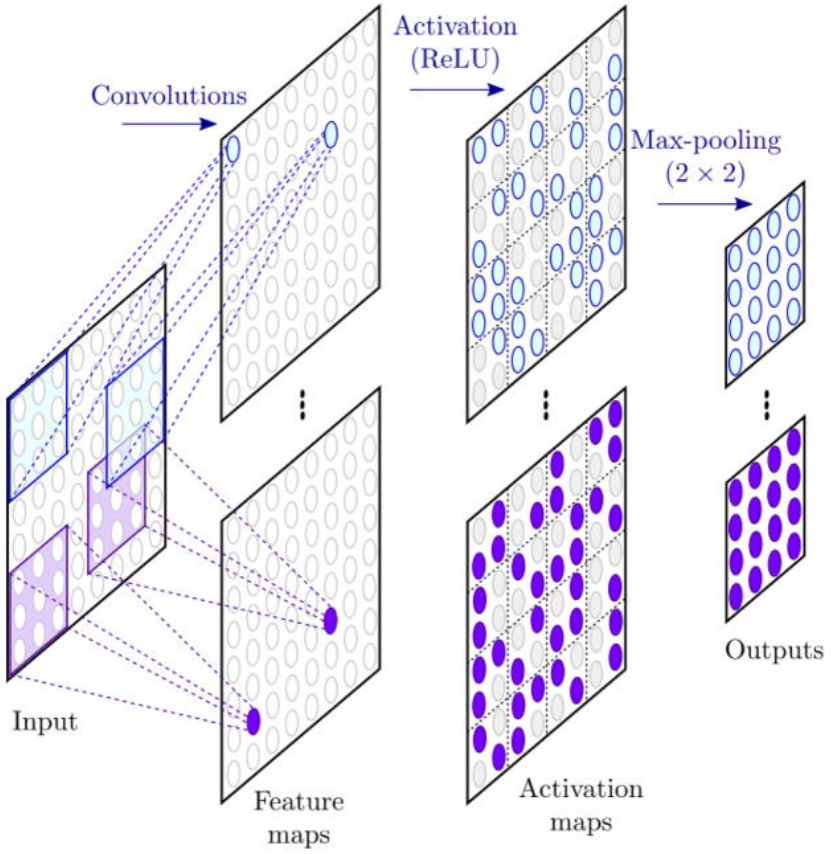


Figure 3.2: A typical sub-model arrangement seen in convolutional neural networks. Two 3 x 3 convolutional kernels (blue and purple) operate over the convolutional layer input. Each feature map has an associated kernel. The ReLU function performs non-linear activation on each extracted map. Finally, a 2 x 2 max-pooling layer halves the output dimensions and encodes a 2 x 2 translational in-variance for selected features in each partition [30].

3.2.1 Convolution layers

Each neuron (or node) in the feature maps receives input only from a restricted subsection of the previous layer ('local connectivity'), known as the receptive field, determined by the size of the convolutional kernel [30]. A fundamental assumption of FCNs is that spatially close input neurons (i.e. input pixel values) have an increased significance for pattern recognition when compared to distant pixels [54]. For instance, a 3 x 3 kernel will convolve around activation values from the previous layer, and each node in the output feature map will have a receptive field size of 9 pixels (3 x 3) corresponding to a location in the input. In the case of CNNs, filter values in each convolutional kernel are the model parameters; optimised to extract features with predictive validity during the back-propagation phase of training [29]. It follows that larger filter sizes output single node values (and hence feature maps) that are representative of a spatially broader subset of information [16]. Each filter in a convolutional layer is associated with a feature representation of the input (known as a feature map) [30]. In contrast to the fully connected perceptron networks described in section 3.1, a single filter convolves with the same kernel weights over the entire input domain - ensuring convolutional in-variance across an individual feature map [29] - hence a specific pattern recognition task can be associated with each kernel in the layer [55]. In addition to making sense conceptually (i.e. edge detection is likely as useful over the entire input as it is on a kernel-sized subset), this 'parameter sharing' has the added advantage of reducing the total number of parameters in a model; reducing both

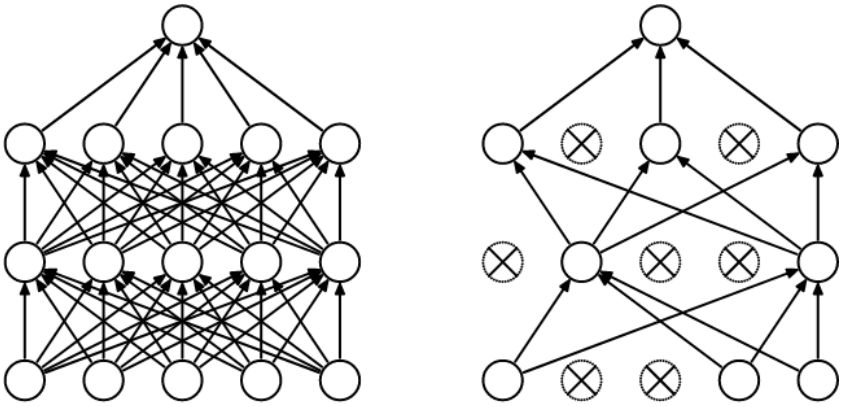
computational complexity and GPU memory requirements [34].

3.2.2 Pooling layers

Pooling is a technique for sub-sampling feature maps in order to decrease resolution; and, to encode translational in-variance to activation values in its receptive field [34]. For instance, a 2 x 2 max-pooling layer outputs the maximum value from each non-overlapping 2 x 2 grid partitioned from its input, halving the total feature map dimensions in the x-y plane, and encoding spatial in-variance with respect to the selected value over the 2 x 2 grid [34]. Downsampling in this way is also possible via a 2D convolution with a stride size of 2. Convolutional pooling adds trainable parameters in the form of a kernel and has the benefit of simplifying the overall network structure due to the linear nature of convolution [56].

3.2.3 Dropout layers

Dropout layers enforce a regularisation constraint by randomly sampling neurons in a network for deactivation; enforcing redundancy in a network, as adjusted architectures process each training batch [34]. Weights are therefore optimised on multiple sub-variations of the complete network, resulting in stochastic averages with values sampled from a broad ensemble of networks [57]. Dropout layers improve generalisation and hence model performance when compared to fixed architectures. We note that dropout is deactivated during post-training inference, and predictions occur via the complete architecture [34].



(a) Standard Neural Net (b) After applying dropout.

Figure 3.3: Multi-layer perceptron with 2 hidden layers. (a) Standard network without dropout applied. (b) Standard network after applying dropout. The dropout technique randomly samples neurons in the network for deactivation, allowing parameter tuning to occur as averages over an ensemble of networks [57].

3.2.4 Batch normalisation layers

Batch normalisation layers are typically placed after activation layers (ordering is a topic of current research and alternative variations exist i.e. [9]) to normalise activated feature maps. Hence, each batch normalisation layer adds two trainable parameters to the model, a mean and a variance value [34]. Although batch normalisation has been shown to accelerate training and improve stability in layer distributions, its effectiveness is poorly understood from a theoretical perspective [38]. Conventional understanding states that batch normalisation penalises internal

co-variance between network layers [12] and reduces effects of the vanishing/exploding gradient problems [58]. Additionally, parameter updates to feature maps in earlier layers are likely to significantly change the feature distribution of deeper layers - compounding internal variance in the optimisation process [38]. Hence, by normalising the output activation map, changes in layer distributions can be stabilised [38]. Conversely, studies have reported minimum improvements to internal co-variance under batch normalisation; and report that performance improvement is due to a smoothing of loss topology, which stabilises the behaviour of gradient descent [38].

3.2.5 Activation functions

In the previous section, we presented the fundamental structure of MLPs and expanded on this by examining CNN layers. However, 'deep' learning is not possible with the typical sigmoid activation function presented in early MLP networks [34] (included in equation 3.3) due to the vanishing gradient problem - whereby multiple derivatives in the back-propagation algorithm with values < 1 cause the loss gradient (equation 3.1) to decay exponentially as a function of the number of layers [34].

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.3}$$

In contrast, convex functions such as the piecewise linear ReLU (Rectified linear unit - equation 3.4) have shown improved performance in deep layer arrangements due to the sparse nature of their activation [59]; and, as non-saturating gradients prevent the vanishing gradient problem [34]. However, contrary to activation function requirements stated in section 3.1, ReLU is neither smooth nor bounded, raising some technical challenges in its use [34]. The unbounded nature of ReLU exposes the network to the well-known 'exploding gradient problem' - as there is no constraint on activation values [60]. In addition, the 'Dying ReLU problem' highlights a limitation associated with zeroing all negative activation inputs [60]. Zeroed activation values may fail to influence the loss function during gradient calculations; hence, this state partially prohibits parameter adjustment during training [60]. Although some recent CNN implementations attempt to offset the Dying ReLU problem via the so-called 'Leaky ReLU' (which maintains a small non-zero gradient for activations < 0 [61]), ReLU is still the default recommendation for most deep neural networks [62].

$$ReLU(x) = max(0, x) \tag{3.4}$$

4 Method

4.1 Datasets

This project focused on building an automated segmentation solution for radiotherapy. Two applications were targeted: A multi-organ segmentation model for pelvic imaging (with patient, bladder and rectum contouring); and a single structure model for vacuum bag contouring in canine imaging. Anonymised pelvic imaging data was provided by Riverina Cancer Care Centre from active prostate cancer RT patients over multiple stages of treatment. Canine imaging data was provided by the Small Animal Specialists Hospital (SASH) and contained variable cancer locations and patient orientations. All input data consisted of raw diagnostic CT images acquired with a 512 x 512 matrix. Pelvic imaging scans were comprised of 1.37 mm x 1.37 mm x 2 mm voxels; while canine imaging scans contained variable spacings across patients, with an average voxel size of 0.85 mm x 0.85 mm x 1.907 mm.

Patient scans were converted from a propriety Monaco format (.WC files) to DICOM files, from which image-structure pairs were extracted, transformed from patient-space to a non-dimensional matrix-space, and saved individually as model input-output arrays for further processing. Initial modelling was attempted with contours extracted on-the-fly from DICOM files; however, this resulted in a significant CPU bottleneck which limited GPU capacity during training. In addition, any advantage of removing the intermediate file processing step (DICOM to array) was made redundant upon determining that significant data cleaning would be required for contour consistency. For instance, vacuum bag segmentation was often incomplete in patient scans, as only clinically relevant locations included full contouring; this may satisfy clinical requirements, however, consistent labels are required for machine learning. The final data pipeline was designed to handle filenames (pointing to arrays) as the primary method of matching an input with the ground truth. From here, an array for each filename was read into memory if it belonged to the current batch; this removed memory constraints on dataset size as only a single batch populated the RAM at each training step.

A total of 15 patients were used for pelvic imaging, corresponding to 1991 total input instances. Data was split at the patient level to enforce independence across training, validation and test datasets. 12 patient scans were used for training, while validation and testing used 2 and 1, respectively. The complete data distribution is provided in Table 4.1. Patient contours were present in all input data, while the bladder was present in 28% of slices, and rectum in 37%. We note the significant pixel-wise class imbalance between structures, as the output space was 512 x 512 x 3 for multi-organ segmentation. Patient pixels corresponded to a total of 5.21% of all output pixel in the data, with 0.081% and 0.021% corresponding to bladder and rectum.

Table 4.1: Data distribution for pelvic imaging.

	Training	Validation	Testing	Total
Images(Patients)	1751(12)	282(2)	138(1)	1991(15)

	Images total (%)	Pixel-image ratio (%)	Pixel-output ratio (%)	Pixels total (%)
Patient	100	15.6	5.21	5.21
Bladder	28.0	0.862	0.287	0.081
Rectum	37.0	0.172	0.057	0.021

For vacuum bag segmentation, a total of 26 patients were used, with 21, 3, and 2 corresponding to training, validation and testing, respectively. Vacuum bag structures were present in 70% of total input images, as seen in Table 4.2.

Table 4.2: Data distribution for canine imaging.

	Training	Validation	Testing	Total
Images(Patients)	1912(21)	340(3)	187(2)	2439(26)

	Images total (%)	Pixel-image ratio (%)	Pixel-output ratio (%)	Pixels total (%)
Vacbag	70.0	13.4	13.4	9.4

Significant data augmentation was used to increase the effective size of our dataset and to regularise over-fitting. Augmentation was performed on-demand for each input-output pair in a batch, and sampled from a random uniform distribution with probability listed below for each type. 50% of total training data was selected for augmentation per epoch. Augmentations included: Left-right image inversion ($P_{val} = 0.5$), random image cropping and resizing ($P_{val} = 0.33$, minimum crop size 500 x 500), elastic deformations ($P_{val} = 0.33$, with (α, σ) pairs selected from (1201, 10), (1501, 12), and (991, 8)), affine transformations ($P_{val} = 0.33$, $\alpha_{max} = 20$), and Gaussian noise ($P_{val} = 0.33$, $\mu = 0$, $\sigma_{max} = 0.3$). Furthermore, all input data (including test data) was normalised with respect to the training and validation dataset distributions prior to augmentation. Randomly sampled transformations are included in Figure 4.1.

4.2 Model architecture

We designed a 2D U-net architecture with 7 levels, consisting of 6 encoding and 6 decoding blocks, outlined in Figure 4.2. The model accepts a full resolution (512 x 512) CT image as input, and outputs selected contours in the original resolution by the use of padded convolutions, each of which is followed by batch normalisation and ReLU activation. Each encoding block performs a repeated sequence of 3 x 3 convolution (increasing feature channels). Extracted feature maps are passed via the skip connection in one pathway, while a 3 x 3 convolution with stride size 2 halves the resolution before passing features to the next encoding block.

Conversely, each decoding block upsamples input via a 3 x 3 2D transposed convolution with stride size of 2. Upsampled feature maps are then concatenated with skip connections. Dropout is selectively performed with a probability value of 20%, before additional 3 x 3 convolution sequences reduce the feature channels.

Finally, multi-organ segmentation can be controlled via the C (Channel - corresponding to the number of segmentations) output variable specified in the last 1 x 1 convolutional layer.

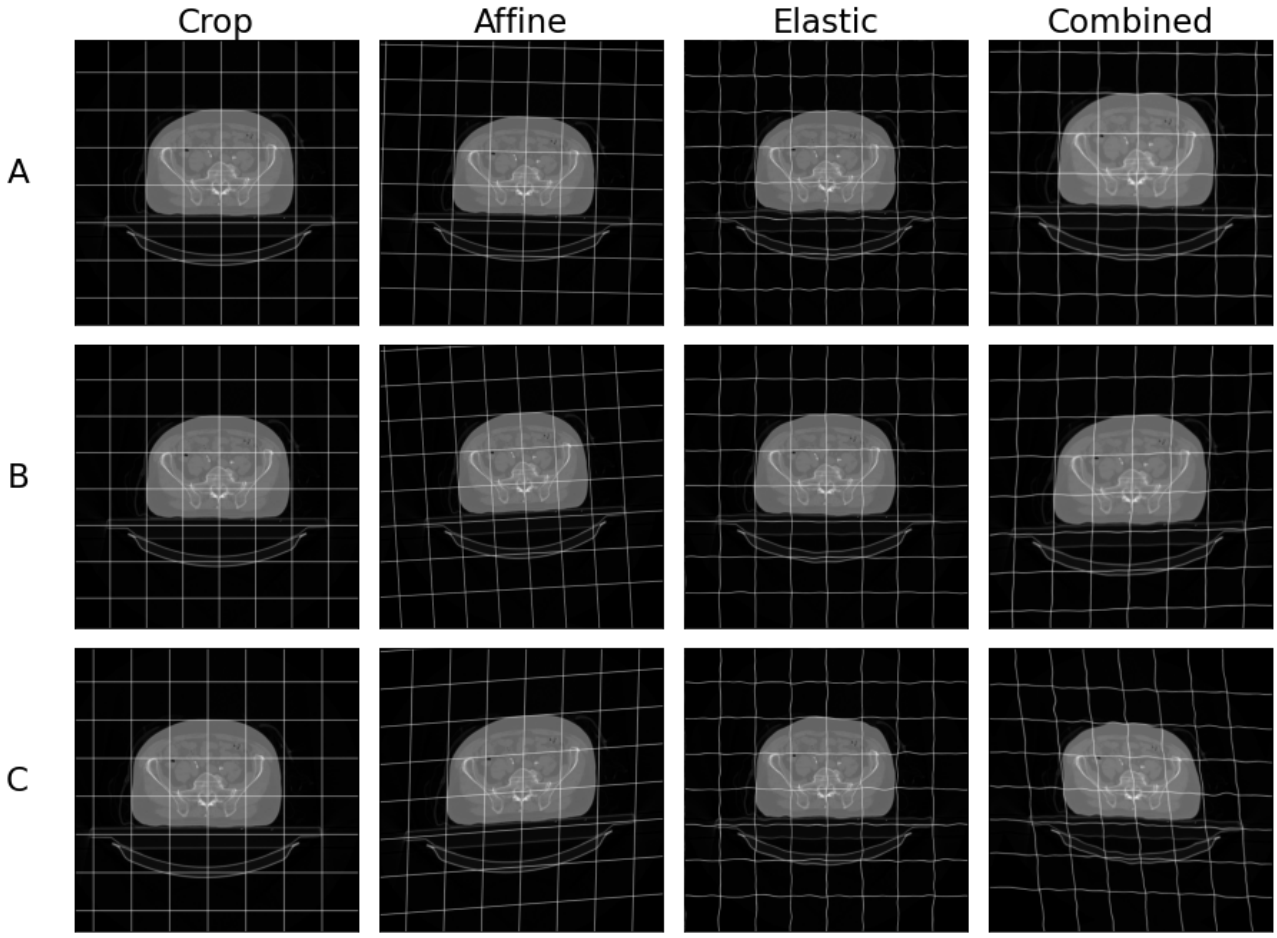


Figure 4.1: Training data augmentation for single input image with random sampling of parameters: image crop and resize, affine transformation, elastic deformation, and combined transformations. Each matching contour set is augmented under an identical transformation. An individual transformation type has $P_{val} = 0.33$ of occurring. Additional augmentations not shown: Left/right inversion and Gaussian noise

A sigmoid activation function is used to account for the non-mutually exclusive nature of pixel-wise binary classification on anatomic structures (i.e. a voxel can belong to multiple structures).

The model was trained using the Adam (Adaptive momentum estimation) optimisation algorithm [63] with an initial learning rate of 10^{-5}, a batch size of 1 for pelvic imaging, and 3 for canine imaging. Model training was scheduled to conclude when validation loss had not improved for a period of 20 epochs. In addition, learning rate decay was triggered by a validation loss plateau period of 3 epochs. Initial model weights were determined via 'He' kernel initialisation), which samples from a zero mean Gaussian distribution with variance $\sigma = \sqrt{2/N}$ (as in Ronneberger et al [11]), where N is the incoming nodes for a single activation (i.e. for n x n convolution over M feature maps, N = n x n x M). In addition, we accelerate training for pelvic imaging by adopting the strategy of Bertels et. al to further initialising model parameters via 3 epochs of training with binary cross entropy.

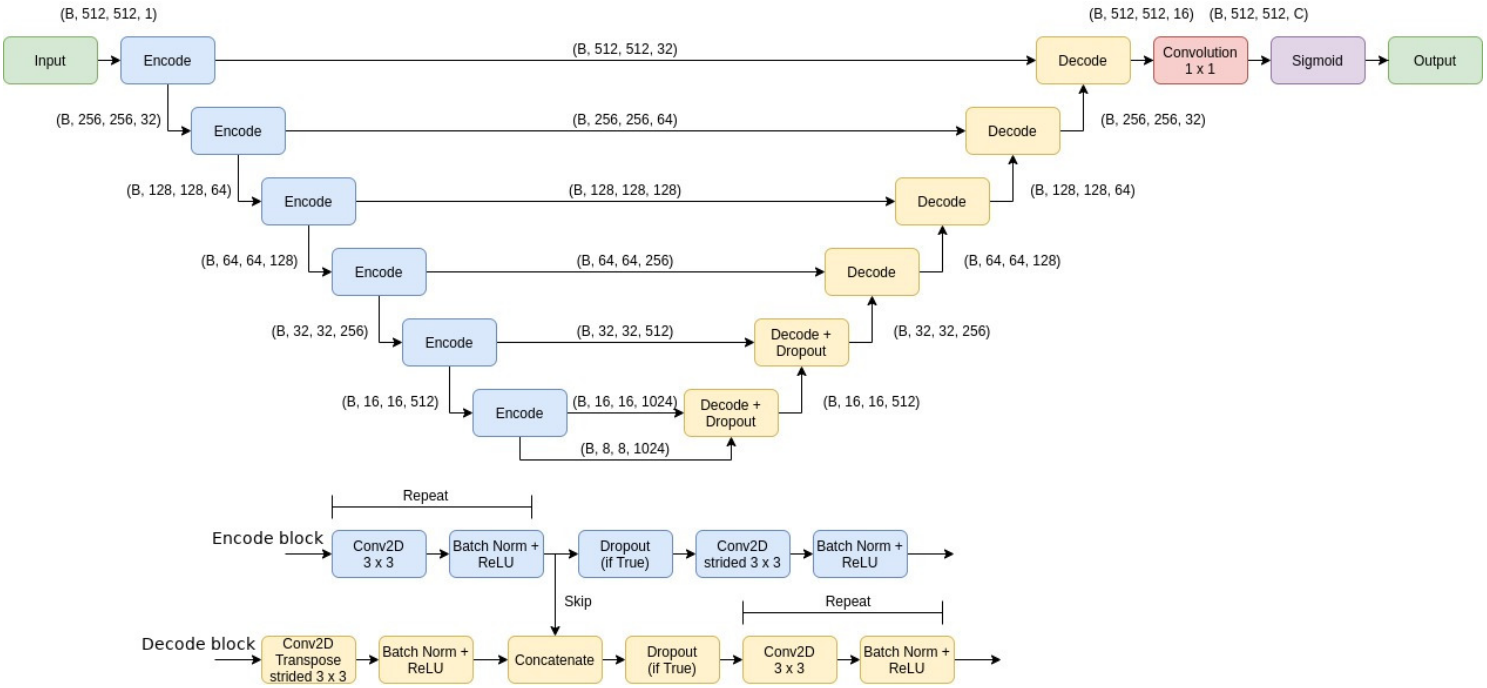


Figure 4.2: Modified 2D U-net architecture: Composed of encoding (blue) and decoding blocks (yellow). Max-Pooling layers replaced by strided convolution. Added batch normalisation and final sigmoid activation.Tensor dimensions (Batch size, X, Y, Channels) are included for each connection. Internal layers of encoding blocks (blue) and decoding blocks (yellow) are included under the high-level overview.

4.3 Loss functions

A total of 5 loss functions were assessed for pelvic imaging: Binary Cross entropy (BCE), soft dice similarity coefficient (soft DSC) [24], weighted soft dice similarity coefficient (w. soft DSC or modified generalised dice loss [47]), a modified combination loss BCE + 2 (w. soft DSC) [45], and focal Tversky loss [10, 36, 48]. In contrast, weighted soft DSC was not attempted for canine imaging due to the single segmentation output.

Calculations for weighted soft DSC were performed via equation 4.1, with component equations presented in 4.2, 4.3, and 4.4. W, I, and U correspond to the weight, intersection and union vectors, indexed with respect to each contour structure by $k \in \{1, ..., n\}$; while ϵ represents a small value > 0 to ensure division is defined. In addition, $|\cdot|$ notation refers to the cardinality of a set (i.e. the total number of pixels in a contour mask - including zeros), while t_k refers to a ground truth array, p_k to the predicted array, and $\odot$ to the Hadamard product. The standard soft DSC can be implemented by setting weights $W = \vec{1}$. Code required for a tensorflow implementation is included in the pymedphys library.

$$w.\,soft\,DSC = \frac{2\,(W \cdot I + \epsilon)}{W \cdot U + \epsilon} \tag{4.1}$$

$$W = (W_1, ..., W_n) \;:\; W_k = \frac{\sum_{k=1}^{n} |t_k|}{t_k + \epsilon} \tag{4.2}$$

$$I = (I_1, ..., I_n) \;:\; I_k = \sum_{i,j=1} (t_k \odot p_k)_{ij} \tag{4.3}$$

$$U = (U_1, ..., U_n) \ : \ U_k = \sum_{i,j=1} (t_k + p_k)_{ij} \tag{4.4}$$

Addition? - Tversky loss formula and parameters used - [48]

Final model performance was evaluated on an independent test dataset via both DSC and sDSC metrics. Organ specific tolerance τ for rectum and bladder contours were taken as the 95th percentile absolute mean surface distance in mm between expert observers from Roach et al. [15]. MSD_{95} values calculated from Roach et al. show organ-specific tolerances of 1.46 mm for the bladder, and 6.99 mm for the rectum.

5 Results and discussion

5.1 Model 1: Pelvic imaging

A total of 5 loss functions were assessed for ability to train a 2D U-net with a small dataset. As seen in Table 5.1, dice similarity coefficient (DSC), precision and sensitivity metrics were recorded on an independent test dataset to measure model generalisability for each loss. Binary cross entropy (BCE) is a pixel-wise similarity measure [24], and achieved the strongest scores with a DSC of 0.996, precision of 0.999, and sensitivity of 0.991. Patient contouring under BCE was excellent, with an average organ specific DSC of 0.996; however, BCE failed to produce any positive predictions for the rectum, and only large bladder examples were identified - likely due to the built-in assumption that classes are balanced [11]. Reports have indicated that BCE has sub-optimal performance on class imbalanced data [45]; however, large bladder examples were contoured accurately (see Figure 5.4).

The standard soft DSC loss function also assumes equally weighted segmentation classes throughout the data [47]; and hence, failed to produce positive predictions for the bladder and rectum. However, patient contouring was again excellent, with mean DSC 0.996. Research has indicated that the DSC is equivalent to the harmonic mean of recall (sensitivity) and precision [64]; and hence, weighs both equally [64] - contributing to the class imbalance problem as the majority of output pixels are negatives. It is not possible to control the trade-off between false-positives and false-negatives in the standard soft DSC loss formulation [45]; as expected, we observe a bias that favours negatives due to their over-representation in model output [45]. To control for class imbalance in pelvic CT imaging (where boundaries between OARs can be poorly defined) we likely require a loss function that enforces a higher penalty for false-negative values, such that target regions are not under-segmented [45]. An additional point of warning to note is that the soft DSC does not include true-negatives in its calculation, hence specificity is not optimised directly [45].

Conversely, the weighted soft DSC loss function (presented in equation 4.1) was the only loss attempted that was able to optimise for all organs in the contour space (see Figures 5.2, 5.3, and 5.5), and was selected as our final model loss. We note that a simplified combination BCE and weighted soft DSC loss was also attempted (see [45]) after experiments revealed BCE performance was superior to the standard soft DSC metric, and contoured larger bladder examples more accurately than the weighted soft DSC. However, only patient contours were produced - reinforcing that scalar selection to optimally balance a linear combination of loss functions is a non-trivial task dependent on the data distribution [24].

Focal Tversky loss (see [36]) exceeded both BCE and the combination loss in sensitivity; however, performed poorly on average volumetric overlap (DSC). Tversky loss successfully identified rectum and bladder contours in almost all cases; however, segmentation masks contained many false-positive results, with additional groupings that neither resembled nor were spatially close to the OAR in question. We note that although Tversky loss aims to improve the trade-off between sensitivity and precision compared to DSC for highly imbalanced data (i.e. by weighting to penalise false-negatives higher than false-positives) [45]; our results in Table 5.1 indicated

sensitivity was lower when compared to BCE and the weighted soft DSC loss. However, Focal Tversky was the only loss function to perform higher in sensitivity than precision - consistent with a higher weighting on false-negatives as described in the literature [36].

Table 5.1: Loss evaluation on independent test dataset for pelvic imaging

Loss	DSC	Precision	Sensitivity
BinaryCrossentropy (BCE)	0.995	0.999	0.991
soft DSC	0.986	0.999	0.972
w. soft DSC	**0.994**	**0.997**	**0.991**
BCE + 2(w. soft DSC)	0.985	0.999	0.971
FocalTversky	0.962	0.941	0.987

Our final model for pelvic imaging was selected at the 140th epoch, under weighted soft DSC. As seen in Figure 5.1, validation loss plateaued at 130 epochs, with DSC value reaching a maximum at 140 epochs. We note that although a smoothed representation of loss for both validation and training data would be monotonically decreasing, this is not the case for DSC, precision, and sensitivity metrics. In section 4.2 we noted the use of BCE weight initialisation; after switching from BCE to weighted soft DSC, there is a significant change in loss topology. Examination of early model predictions after BCE initialisation indicated a change in state from placing minimal significance on bladder and rectum cases, to now having gradients dominated by changes to these organs under weighted soft DSC. It is likely that feature representation were subsequently perturbed significantly to minimise the new loss - temporarily decreasing DSC and precision. A similar pattern emerged for DSC value under Focal Tversky; however, the initial DSC value is comparatively lower than under weighted soft DSC - indicating that convergence after 3 epochs of BCE is likely sensitive to randomly sampled weight initialisation [11]. Total inference time per patient was approximately 3 seconds.

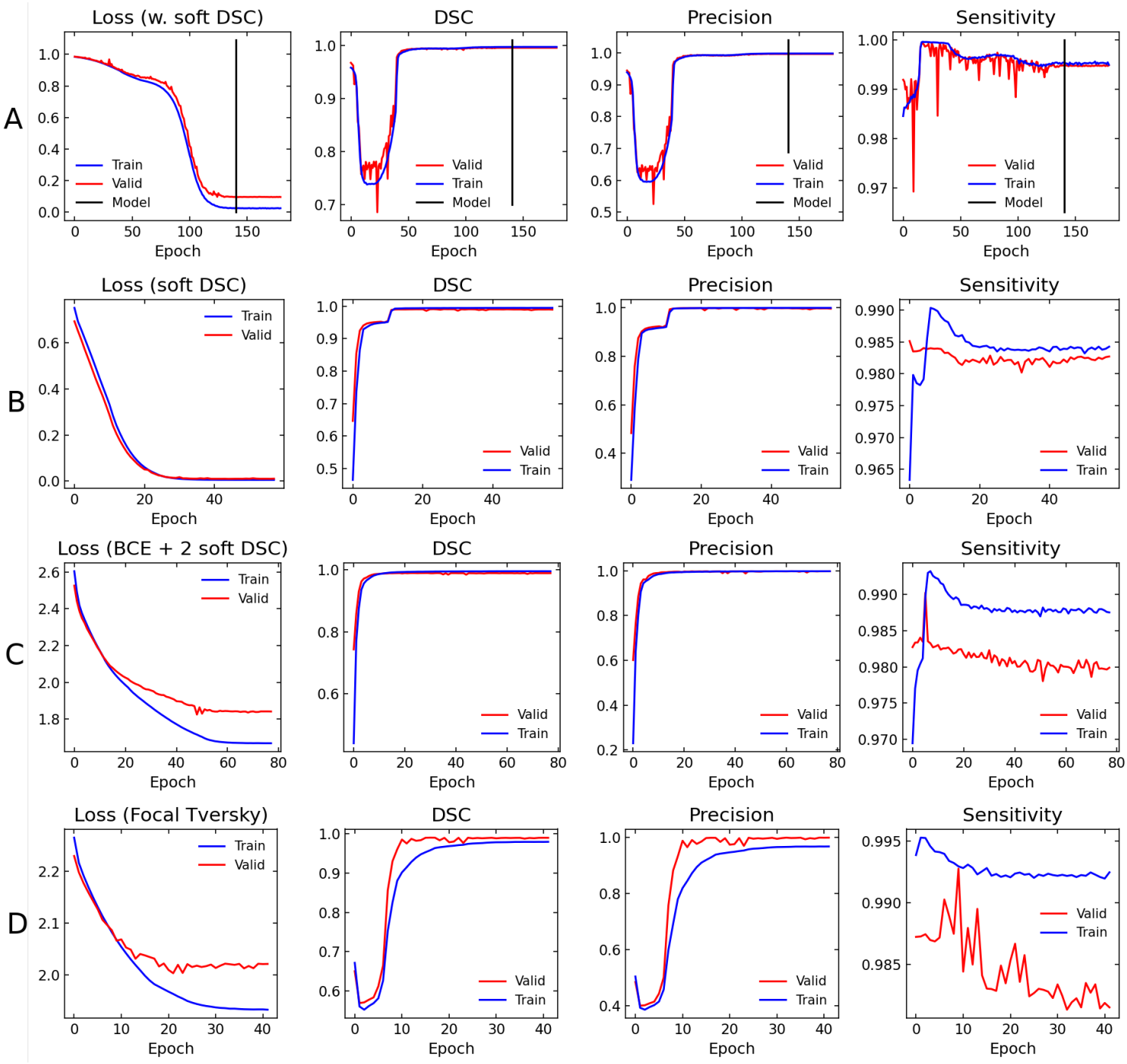


Figure 5.1: **A)** Model training metrics for pelvic imaging via weighted soft DSC loss (w. soft DSC). Final model selected at epoch 140 due to validation loss plateau. Metrics begin post binary cross entropy (BCE) weight initialisation (3 epochs). Training time of 9 hours.
B) Soft dice similarity coefficient (soft DSC) loss
C) Combination binary cross entropy (BCE) and weighted soft dice similarity coefficient (w. soft DSC) loss
D) Focal Tversky loss

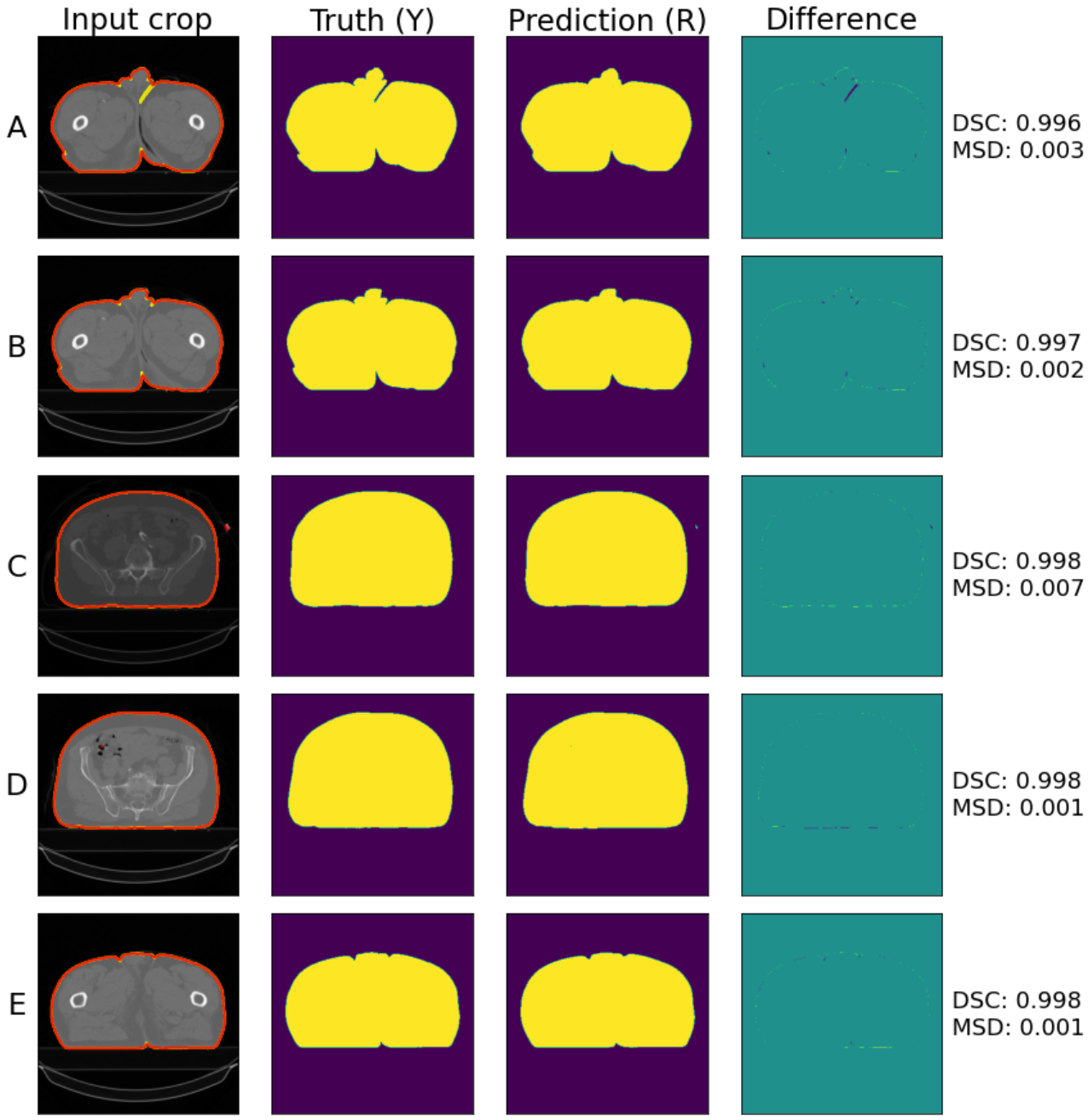


Figure 5.2: Representative output for patient. Model 1 - trained via weighted soft dice (w. soft DSC) loss - 140 epochs. Truth contour (yellow), prediction contour (red). Mean surface distance (MSD) mm.

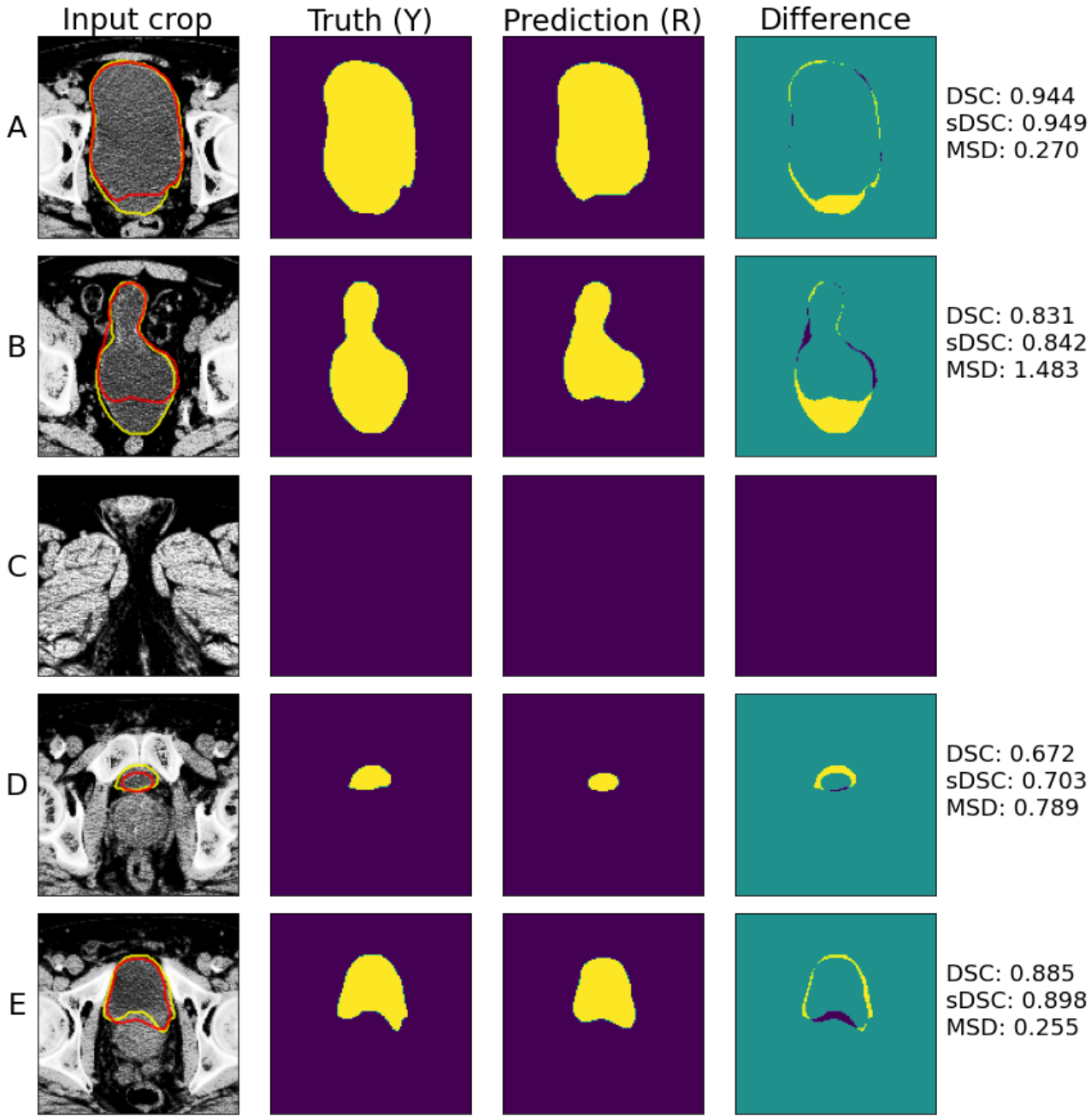


Figure 5.3: Representative output for bladder: Model 1 - trained via weighted soft dice (w. soft DSC) loss - 140 epochs. Truth contour (yellow), prediction contour (red). Mean surface distance (MSD) in mm. sDSC [5] calculated at an organ specific tolerance of $\tau = 1.46$ mm, the 95th percentile mean surface distance between expert observers [15].

In Figure 5.3, weighted soft DSC underestimated the posterior aspect of larger bladder examples - a recurrent limitation of the model. Figure 5.3 C) shows the model was correctly able to identify when contours were not present, indicating the strong negative predictive validity of the model - quantified by the high sensitivity score recorded on Table 5.3.

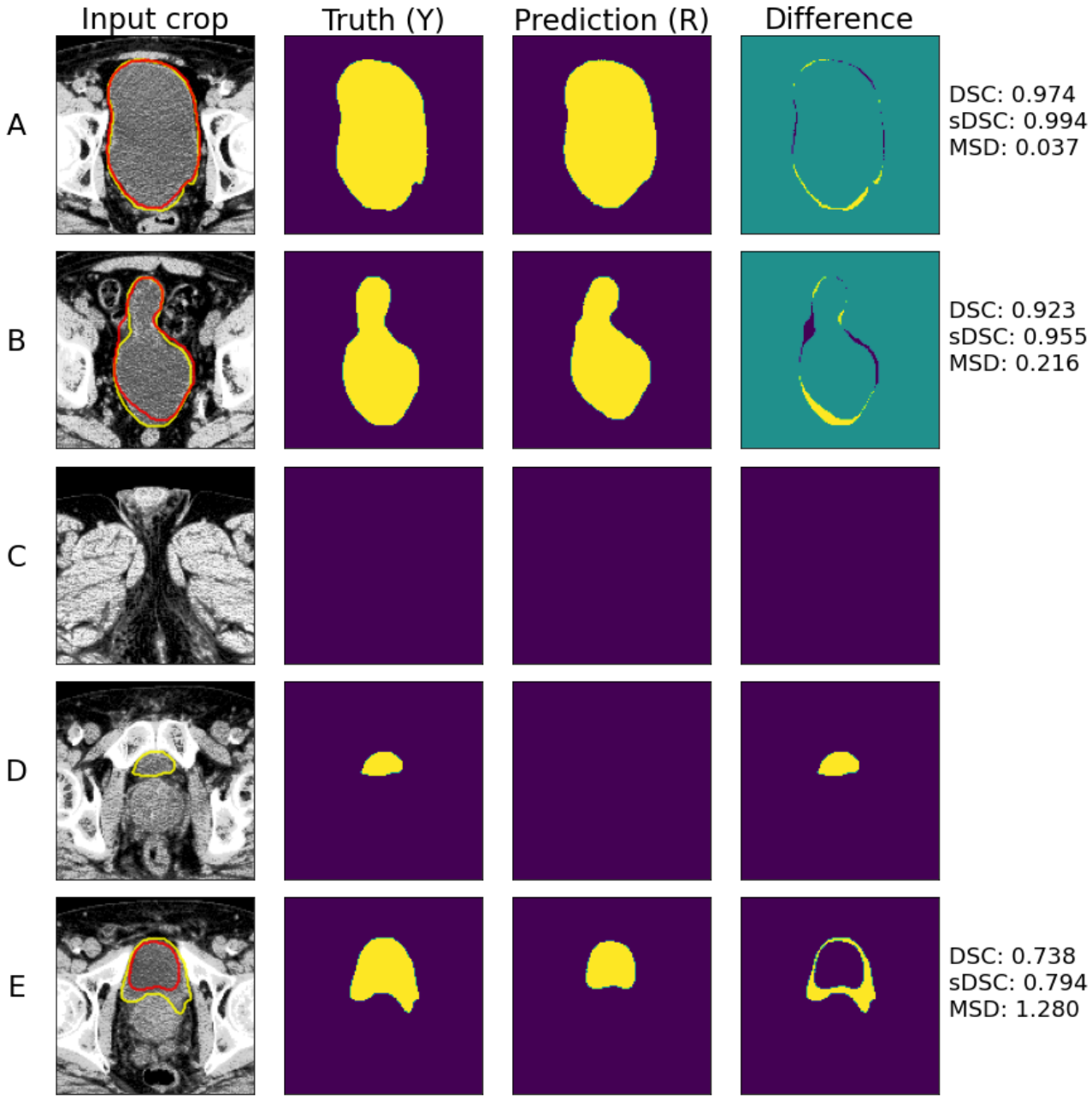


Figure 5.4: Representative model output for bladder: Trained via binary cross entropy loss - 78 epochs. Truth contour (yellow), prediction contour (red). Mean surface distance (MSD) in mm. sDSC [5] calculated at an organ specific tolerance of $\tau = 1.46$ mm, the 95th percentile mean surface distance between expert observers [15].

Model 1 under weighted soft DSC was unable to identify rectum contours containing hollow regions (see Figure 5.5 D). We suspect further training on a distribution of similar cases may improve performance. All other cases in the test dataset were correctly identified. As seen in Figure 5.5 and Table 5.3, DSC values for rectum contours were lower on average when compared to the bladder - consistent with expert IOV [**CITATION**]and other models in the literature [**CITATION**]. However, a higher surface dice similarity coefficient (sDSC) indicated that the degree of correction required for rectum contours was lower than corrections required for bladder contours.

Cite

Cite

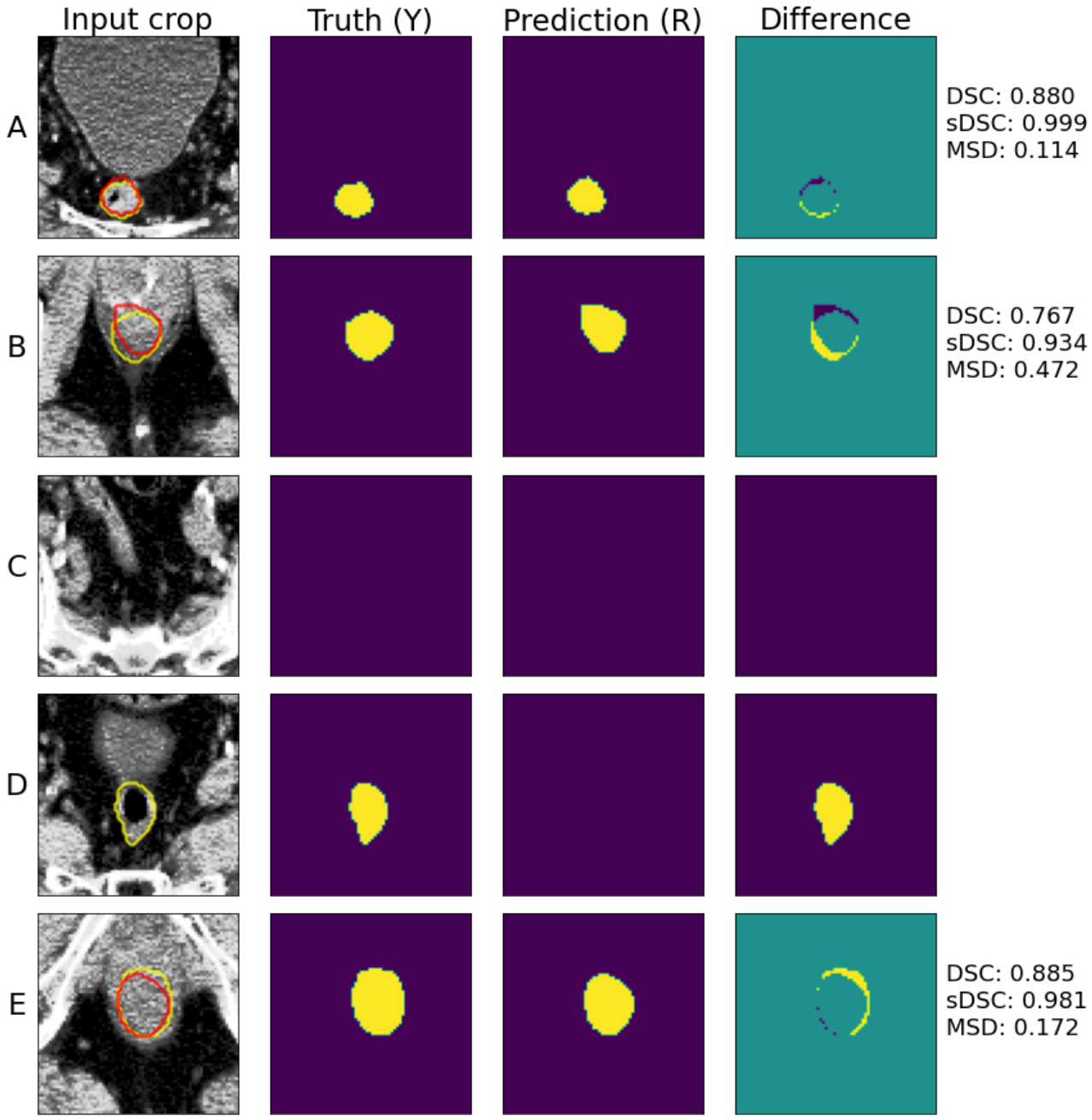


Figure 5.5: Representative output for rectum: Model 1 - trained via weighted soft dice (w. soft DSC) loss - 140 epochs. Truth contour (yellow), prediction contour (red). Mean surface distance (MSD) in mm. sDSC [5] calculated at an organ specific tolerance of $\tau = 6.99$ mm, the 95th percentile mean surface distance between expert observers [15].

5.2 Model 2: Canine imaging

3 loss functions were attempted for vacuum bag segmentation in canine imaging, as seen in Table 5.2. Soft DSC outperformed both BCE and focal Tversky on DSC and precision values. Focal Tversky had the highest sensitivity (0.969) as expected [36], with BCE second (0.954). The final model was selected at 100 epochs under soft DSC loss. Representative model output is presented in Figure 5.7. Soft DSC loss showed excellent agreement with ground truth vacuum bag contours - and was able to handle both negative (E) and small contour (B) examples in the test dataset (see Figure 5.7).

Table 5.2: Loss evaluation on independent test dataset for canine imaging

Loss	DSC	Precision	Sensitivity
BinaryCrossentropy	0.901	0.935	0.954
soft DSC	**0.952**	**0.953**	**0.953**
FocalTversky	0.930	0.906	0.969

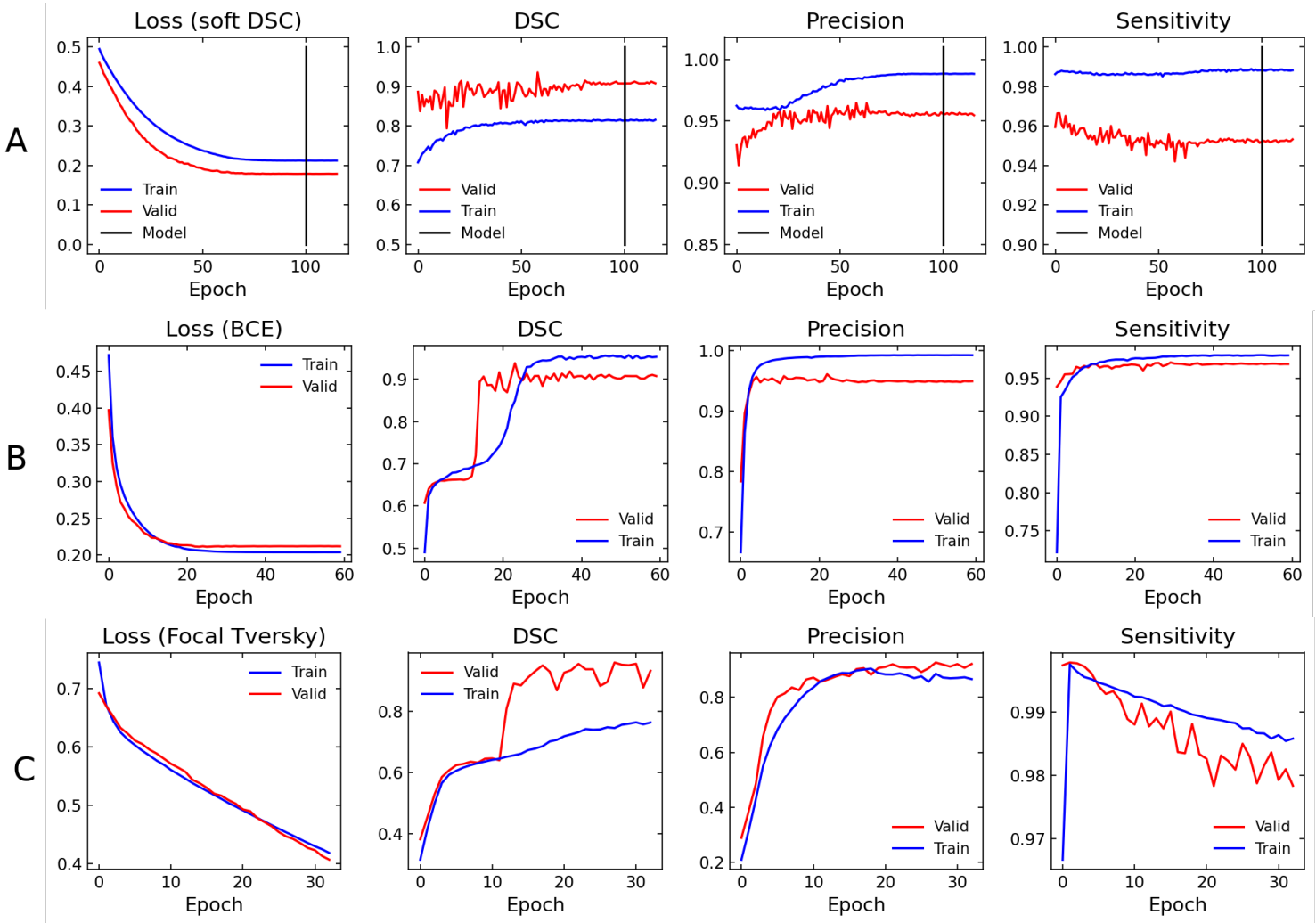


Figure 5.6: **A)** Model training metrics for canine imaging via soft dice similarity coefficient (soft DSC) loss. Final model selected at epoch 100 due to validation loss plateau. Training time 6 hours.
B) Binary cross entropy (BCE) loss
C) Focal Tversky loss

Typically, we expect training metric values to overstate a model's predictive capacity, as parameter values are updated to fit the training data distribution. However, in Figure 5.6 we see that at many stages, validation loss was lower than training loss. The literature states two possible contributing factors: 1) Dropout layers regularise only on the training data, and hence the full architecture is only available for inference on the validation and testing sets [57]. Additionally, batch normalisation parameters are tuned to normalise activations on the test dataset - and are fixed during validation and testing [**CITATION**]. 2) Due to the small dataset used in this study, variation in the validation set may itself be small compared to the training data; hence, if the validation distribution is centred about the mean of the training data, the validation

Cite

dataset would be relatively 'easier' to infer [**CITATION**]. We note that although Tversky loss continued to decrease for both the training and validation sets over the epochs tested, a DSC validation plateau (see oscillation in Figure 5.6 C)) triggered early stopping.

Cite

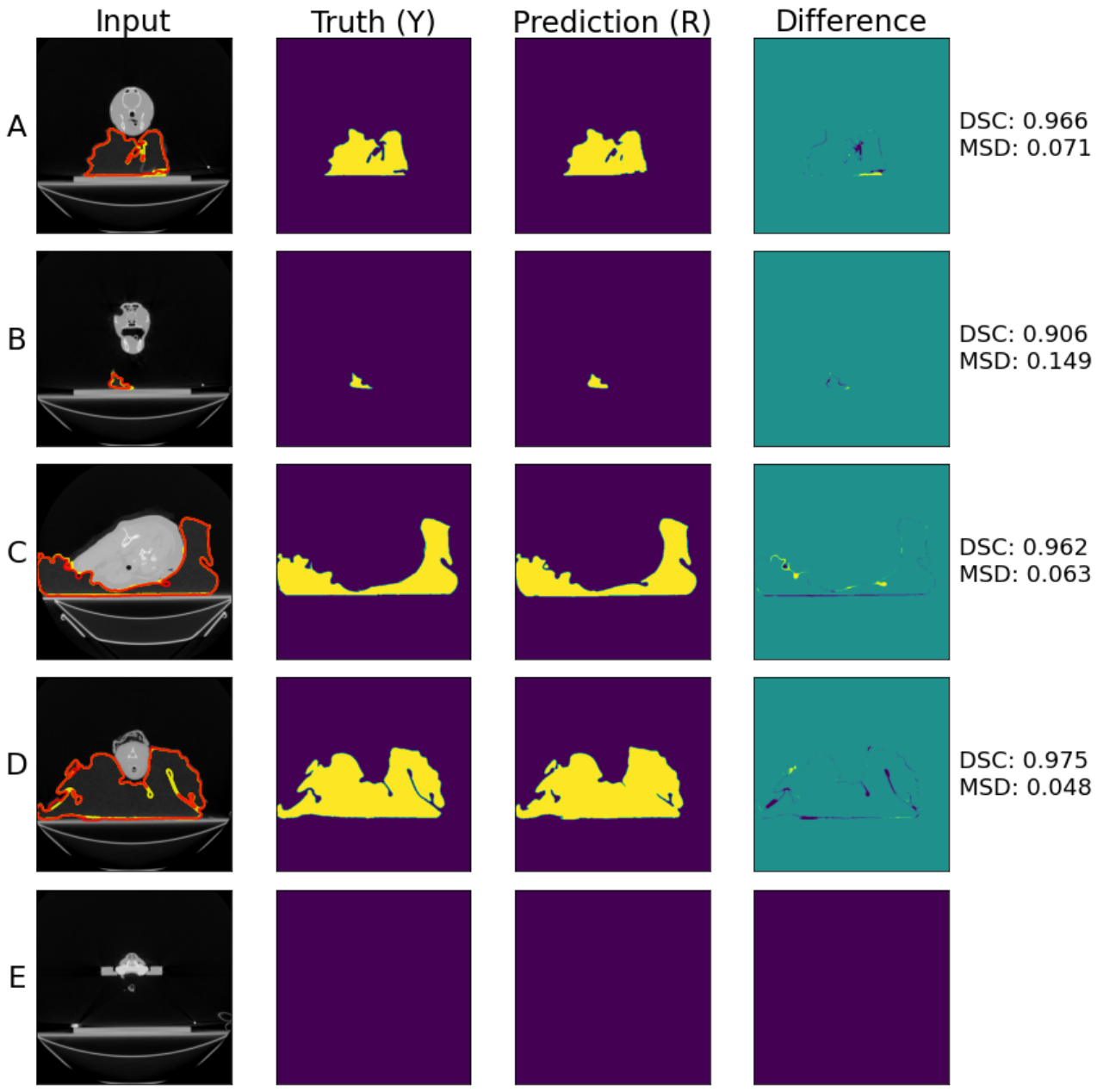


Figure 5.7: Representative output for vacuum bag: Model 2 - trained soft DSC loss. Truth contour (yellow), prediction contour (red). Mean surface distance (MSD) mm.

5.3 Clinical relevance

The pelvic imaging model showed excellent agreement with observers for patient contours, with an average DSC value of 0.998, and a mean surface distance (MSD) of < 0.1 mm, as seen in Table 5.3. Bladder contours had an average DSC score of 0.864 with a standard deviation (STD) of 0.221. Additionally, average bladder MSD was measured to be 1.075 mm with an STD of 2.986 mm. Average volumetric rectum agreement was considerably lower, with an average DSC of 0.670 and STD 0.121, and a MSD of 1.120 mm with STD 2.143 mm. In comparison,

organ specific tolerances used for these OARs (95th percentile MSD between experts - i.e. top 95% expert variance) were determined to be 1.46 mm and 6.99 mm for the bladder and rectum, respectively [15, 5]. Large variances in bladder MSD corresponded to predictions that under-segmented the posterior aspect of the bladder (as seen in Figure 5.3 A-B) However, the vast majority of bladder surfaces were contoured correctly within expert IOV - with a mean sDSC 0.876(0.117). In comparison, an even larger proportion of rectum surface points did not need correction to be within expert IOV, with a mean sDSC of 0.922(0.138) recorded. More work is required to correlate sDSC values with time required for contour correction [5, 8].

Literature reports that clinically acceptable agreement between expert observers is DSC ≥ 0.7 for the bladder and rectum [15]. However, experts have been reported to acheive similarities much higher than this, DSC 0.93 ± 0.03, MSD 0.99(0.30) mm for the bladder and DSC 0.81 ± 0.07, MSD 2.862(2.066) mm for the rectum [15]. These finding indicate that although rectum contours produced by model 1 may be clinically acceptable, model performance falls short of expert IOV without contour correction.

State-of-the-art U-net implementations have recently been able to achieve DSC values of 0.95 and 0.92 for the bladder and rectum, respectively [9]. However, 85 CT patient scans were included in this dataset, compared to the 16 included in our study. We suspect that increasing the number of patients in our study would lead to higher generalisability in model performance (reported to scale logarithmically with dataset size [16]) as well as provide a broader validation and test distribution - improving the reliability and robustness of performance metrics [11].

Model 2 produced vacuum bag contours with an average DSC of 0.952 and MSD of 0.175 mm with STD of 0.275 mm. The vacuum bag material has an electron density of approximately 0.1% of water [**CITATION**]; hence, the 95th percentile vacuum bag MSD (0.726 mm) corresponds to a negligible shift is dose distribution under contours produced by this model - and a potential time saving of 30 minutes per patient. Dose shift has yet to be assessed in clinic - however, we suggest that comparing dose volume histograms between model and expert contours under an identical treatment plan could provide a quantitative measure of clinical acceptability. If acceptance testing validates vacuum bag segmentation, model 2 has the potential to save approximately 30 minutes in treatment planning time per patients [**CITATION**].

Cite

Cite

Table 5.3: Organ specific evaluation for proposed models on independent test dataset

Organ: Mean(Std)	sDSC (τ)	DSC	MSD (mm)	Sensitivity	Specificity
Pelvic imaging					
Patient		0.998(0.001)	0.002(0.005)	0.997	0.999
Bladder (τ 1.46 mm) [2]	0.876(0.177)	0.864(0.221)	1.075(2.986)	0.786	0.999
Rectum (τ 6.99 mm) [2]	0.922(0.138)	0.670(0.121)	1.120(2.143)	0.619	0.999
Average		0.994(0.153)	0.409(1.604)	0.991	0.999
Canine imaging					
Vacbag		0.952(0.001)	0.176(0.275)	0.953	0.995

Addition? Further discuss dose implications. Yaser wanted to see this but i feel like this is tangential to the main points of this report. To assess properly would be a project in itself - its also not discussed in ML lit.

5.4 Limitations and future work

A common challenge in deep-learning applications to medical imaging is small data set sizes [11]. Limited data reduces model generalisability [65]. With state-of-the-art implementations using up to 1000 patients per study [5], we expect our results for bladder and rectum segmentation could improve significantly with the broader distribution provided by a larger dataset.

Additionally, organ-specific tolerances used in this study were acquired by Roach et al. from 15 expert observers (9 of which were radiation oncologists) averaged over a cohort of 5 patients [15]. Reliability in these values could be improved by surveying a broader range of experts and patients.

To best of our knowledge, only 1 attempt to correlate sDSC with the time required for contour correction currently exists in the literature. Therefore, more work is required to assess the utility of sDSC in a clinical workflow. Alternative surface-based metrics have also been presented; for example, the estimated added path length in Vaassen et al. (seen in Figure 2.2). However, no study currently exists comparing sDSC and estimated added path length under IOV tolerances. Additionally, there is an opportunity to investigate the barriers and limitations in designing a soft surrogate sDSC metric that can be optimised directly during training.

Furthermore, our ultimate responsibility lies in improving patient outcomes. Hence, there is an opportunity to correlate DSC and sDSC performance with changes to dose distribution when compared with plans developed under expert contouring. A potential advantage of sDSC compared to DSC is the stronger correlation with the time required for contour correction - however, correlation with dose shift may also be an important clinical indicator.

Current studies are researching potential improvements in medical imaging segmentation under 3D U-net models. Although this study focused on a 2D implementation due to the clinical barriers inherent in 3D models, more work is still required to quantify the potential for performance improvement.

6 Conclusion

In this study, we attempted 2D U-net architecture with a small dataset and image augmentation, over a variety of standard loss functions used in semantic segmentation tasks. Two models were developed: Model 1 aimed to contour patient, bladder, and rectum structure in pelvic CT images, to provide a QA tool for background monitoring of IOV in RT. Additionally, we provided surface dice similarity coefficients for the bladder and rectum contours, with organ-specific tolerances at the 95th percentile mean surface distance between expert observers. Model 2 aimed to automate the contouring of vacuum bags in canine imaging for RT, a time-consuming structure that is delineated manually at SASH veterinary clinic.

Weighted soft DSC loss was selected for the pelvic imaging model as it was the only loss that overcame the class imbalance in our data to optimise for all OARs in the model output. Patient contours were excellent, with a DSC of 0.998. We suspect more data will be required to improve bladder and rectum segmentation for use as a QA tool, with DSCs of 0.860 and 0.670, respectively. State-of-the-art implementations employ dataset on the order of 1000 patients, compared to the 16 included in our study. However, the surface dice similarity coefficient indicated that the majority of bladder and rectum surfaces were within expert IOV, with sDSCs of 0.876 and 0.922, respectively - indicating the potential for time saving in contour correction.

Soft DSC loss was selected for the canine imaging model, which produced (likely) acceptable vacuum bag contouring with a DSC of 0.952. As this model has the potential to save 30 minutes of planning time per patient, further work will involve clinical acceptance testing and implementation.

References

[1] Freddie Bray et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer Journal for Clinicians* 68.6 (2018), pp. 394–424. DOI: 10.3322/caac.21492. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21492. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21492.

[2] Michael B. Barton et al. "Estimating the demand for radiotherapy from the evidence: A review of changes from 2003 to 2012". In: *Radiotherapy and Oncology* 112.1 (July 2014), pp. 140–144. DOI: 10.1016/j.radonc.2014.03.024. URL: https://doi.org/10.1016/j.radonc.2014.03.024.

[3] Constantin Dreher et al. "Effective radiotherapeutic treatment intensification in patients with pancreatic cancer: Higher doses alone, Higher RBE or both?" In: *Radiation Oncology* 12 (Dec. 2017). DOI: 10.1186/s13014-017-0945-2.

[4] Shalini K Vinod et al. "A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology". In: *Journal of Medical Imaging and Radiation Oncology* 60.3 (2016), pp. 393–406. DOI: 10.1111/1754-9485.12462.

[5] Stanislav Nikolov et al. *Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy*. 2018. arXiv: 1809.04430 [cs.CV].

[6] Zhikai Liu et al. "Segmentation of organs-at-risk in cervical cancer CT images with a convolutional neural network". In: *Physica Medica* 69 (Jan. 2020), pp. 184–191. ISSN: 1120-1797. DOI: 10.1016/j.ejmp.2019.12.008. URL: http://dx.doi.org/10.1016/j.ejmp.2019.12.008.

[7] Jan Schreier et al. "Clinical evaluation of a full-image deep segmentation algorithm for the male pelvis on cone-beam CT and CT". In: *Radiotherapy and Oncology* 145 (Apr. 2020), pp. 1–6. ISSN: 0167-8140. DOI: 10.1016/j.radonc.2019.11.021. URL: http://dx.doi.org/10.1016/j.radonc.2019.11.021.

[8] Femke Vaassen et al. "Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy". In: *Physics and Imaging in Radiation Oncology* 13 (Jan. 2020), pp. 1–6. ISSN: 2405-6316. DOI: 10.1016/j.phro.2019.12.001. URL: http://dx.doi.org/10.1016/j.phro.2019.12.001.

[9] Samaneh Kazemifar et al. "Segmentation of the prostate and organs at risk in male pelvic CT images using deep learning". In: *Biomedical Physics and Engineering Express* 4.5 (July 2018), p. 055003. ISSN: 2057-1976. DOI: 10.1088/2057-1976/aad100. URL: http://dx.doi.org/10.1088/2057-1976/aad100.

[10] Wentao Zhu et al. "AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy". In: *Medical Physics* 46.2 (Dec. 2018), pp. 576–589. ISSN: 0094-2405. DOI: 10.1002/mp.13300. URL: http://dx.doi.org/10.1002/mp.13300.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[12] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG].

[13] Kaiming He et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. ICCV '15. USA: IEEE Computer Society, 2015, pp. 1026–1034. ISBN: 9781467383912. DOI: 10.1109/ICCV.2015.123. URL: https://doi.org/10.1109/ICCV.2015.123.

[14] Ren Wu et al. "Deep Image: Scaling up Image Recognition". In: (Jan. 2015). DOI: 10.1038/nature0693.

[15] Dale Roach et al. "Multi-observer contouring of male pelvic anatomy: Highly variable agreement across conventional and emerging structures of interest". In: *Journal of Medical Imaging and Radiation Oncology* 63.2 (2019), pp. 264–271. DOI: 10.1111/1754-9485.12844.

[16] Takafumi Nemoto et al. "Efficacy evaluation of 2D, 3D U-Net semantic segmentation and atlas-based segmentation of normal lungs excluding the trachea and main bronchi". In: *Journal of Radiation Research* 61.2 (Feb. 2020), pp. 257–264. ISSN: 1349-9157. DOI: 10.1093/jrr/rrz086.

[17] L. A. Kachnic et al. "RTOG 0529: a phase 2 evaluation of dose-painted intensity modulated radiation therapy in combination with 5-fluorouracil and mitomycin-C for the reduction of acute morbidity in carcinoma of the anal canal". In: *Int. J. Radiat. Oncol. Biol. Phys.* 86.1 (May 2013), pp. 27–33.

[18] L. J. Peters et al. "Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02". In: *J. Clin. Oncol.* 28.18 (June 2010), pp. 2996–3001.

[19] R. Murakami et al. "Interobserver and Intraobserver Variability in Image Registration for Image Guided Radiation Therapy". In: *International Journal of Radiation Oncology*Biology*Physics* 87 (Oct. 2013), S695. DOI: 10.1016/j.ijrobp.2013.06.1843.

[20] Lee R. Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (1945), pp. 297–302. ISSN: 00129658, 19399170. URL: http://www.jstor.org/stable/1932409.

[21] G. Sharp et al. "Vision 20/20: perspectives on automated image segmentation for radiotherapy". In: *Med Phys* 41.5 (May 2014), p. 050902.

[22] A. C. Riegel et al. "Deformable image registration and interobserver variation in contour propagation for radiation therapy planning". In: *J Appl Clin Med Phys* 17.3 (May 2016), pp. 347–357.

[23] Abdel Aziz Taha and Allan Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool". In: *BMC Medical Imaging* 15.1 (Aug. 2015). ISSN: 1471-2342. DOI: 10.1186/s12880-015-0068-x. URL: http://dx.doi.org/10.1186/s12880-015-0068-x.

[24] Jeroen Bertels et al. "Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory and Practice". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Springer International Publishing, 2019, pp. 92–100. ISBN: 978-3-030-32245-8.

[25] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer New York, 2000. DOI: 10.1007/978-1-4757-3264-1. URL: https://doi.org/10.1007/978-1-4757-3264-1.

[26] Femke Vaassen et al. "Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy". In: *Physics and Imaging in Radiation Oncology* 13 (2020), pp. 1–6. ISSN: 2405-6316. DOI: https://doi.org/10.1016/j.phro.2019.12.001. URL: http://www.sciencedirect.com/science/article/pii/S2405631619300636.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. *Fully Convolutional Networks for Semantic Segmentation*. 2014. arXiv: 1411.4038 [cs.CV].

[28] M. Jorge Cardoso et al., eds. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer International Publishing, 2017. DOI: 10.1007/978-3-319-67558-9. URL: https://doi.org/10.1007/978-3-319-67558-9.

[29] Andreas Maier et al. "A gentle introduction to deep learning in medical image processing". In: *Zeitschrift für Medizinische Physik* 29 (May 2019). DOI: 10.1016/j.zemedi.2018.12.003.

[30] Mohammad Hesam Hesamian et al. "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges". In: *Journal of Digital Imaging* 32 (May 2019). DOI: 10.1007/s10278-019-00227-x.

[31] Christian Szegedy et al. *Rethinking the Inception Architecture for Computer Vision*. 2015. arXiv: 1512.00567 [cs.CV].

[32] Chigozie Nwankpa et al. *Activation Functions: Comparison of trends in Practice and Research for Deep Learning*. 2018. arXiv: 1811.03378 [cs.LG].

[33] Guifang Lin and Wei Shen. "Research on convolutional neural network based on improved Relu piecewise activation function". In: *Procedia Computer Science* 131 (2018). Recent Advancement in Information and Communication Technology: pp. 977–984. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2018.04.239. URL: http://www.sciencedirect.com/science/article/pii/S1877050918306197.

[34] Alexander Selvikvåg Lundervold and Arvid Lundervold. "An overview of deep learning in medical imaging focusing on MRI". In: *Zeitschrift für Medizinische Physik* 29.2 (2019). Special Issue: Deep Learning in Medical Physics, pp. 102–127. ISSN: 0939-3889. DOI: https://doi.org/10.1016/j.zemedi.2018.11.002. URL: http://www.sciencedirect.com/science/article/pii/S0939388918301181.

[35] Alexey Dosovitskiy et al. "Discriminative Unsupervised Feature Learning with Convolutional Neural Networks". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 766–774.

[36] Naimul Mefraz Khan, Nabila Abraham, and Ling Guan. "Machine Learning on Biomedical Images: Interactive Learning, Transfer Learning, Class Imbalance, and Beyond". In: *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (2019), pp. 85–90.

[37] Özgün Çiçek et al. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation". In: *Lecture Notes in Computer Science* (2016), pp. 424–432. ISSN: 1611-3349. DOI: 10.1007/978-3-319-46723-8_49. URL: http://dx.doi.org/10.1007/978-3-319-46723-8_49.

[38] Shibani Santurkar et al. *How Does Batch Normalization Help Optimization?* 2018. arXiv: 1805.11604 [stat.ML].

[39] Carlos E. Cardenas et al. "Advances in Auto-Segmentation." In: *Seminars in radiation oncology* 29 3 (2019), pp. 185–197.

[40] Oscar Acosta et al. "Multi-atlas-based Segmentation Of Pelvic Structures From CT Scans For Planning In Prostate Cancer Radiotherapy". In: *Abdomen and Thoracic Imaging*. Ed. by Ayman S. El-Baz; Luca Saba; Jasjit Suri. Springer US, Nov. 2013, pp. 623–656. DOI: 10.1007/978-1-4614-8498-1_24. URL: https://www.hal.inserm.fr/inserm-00910761.

[41] M. Ayadi et al. "Evaluation of ABASTM : multi-center study in the case of prostate cancer". In: *Physica Medica* 27 (2011). 50èmes Journées Scientifiques de la Société Française de Physique Médicale, Nantes Cité des Congrès, 8-10 Juin 2011, S14–S15. ISSN: 1120-1797. DOI: https://doi.org/10.1016/j.ejmp.2011.06.032. URL: http://www.sciencedirect.com/science/article/pii/S1120179711000779.

[42] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[43] Anjali Balagopal et al. "Fully automated organ segmentation in male pelvic CT images". In: *Physics in Medicine & Biology* 63.24 (Dec. 2018), p. 245015. DOI: 10.1088/1361-6560/aaf11c.

[44] Jordan Wong et al. "Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning". In: *Radiotherapy and Oncology* 144 (Mar. 2020), pp. 152–158. ISSN: 0167-8140. DOI: 10.1016/j.radonc.2019.10.019. URL: http://dx.doi.org/10.1016/j.radonc.2019.10.019.

[45] Saeid Asgari Taghanaki et al. *Combo Loss: Handling Input and Output Imbalance in Multi-Organ Segmentation*. 2018. arXiv: 1805.02798 [cs.CV].

[46] Patrick Ferdinand Christ et al. *Automatic Liver and Tumor Segmentation of CT and MRI Volumes using Cascaded Fully Convolutional Neural Networks*. 2017. arXiv: 1702.05970 [cs.CV].

[47] Carole H. Sudre et al. "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations". In: *Lecture Notes in Computer Science* (2017), pp. 240–248. ISSN: 1611-3349. DOI: 10.1007/978-3-319-67558-9_28. URL: http://dx.doi.org/10.1007/978-3-319-67558-9_28.

[48] Nabila Abraham and Naimul Mefraz Khan. *A Novel Focal Tversky loss function with improved Attention U-Net for lesion segmentation*. 2018. arXiv: 1810.07842 [cs.CV].

[49] Rosangela Cintra and Haroldo Campos Velho. "Data Assimilation by Artificial Neural Networks for an Atmospheric General Circulation Model". In: Feb. 2018. DOI: 10.5772/intechopen.70791.

[50] F. Rosenblatt. *The Perceptron, a Perceiving and Recognizing Automaton Project Para*. Report: Cornell Aeronautical Laboratory. Cornell Aeronautical Laboratory, 1957. URL: https://books.google.com.au/books?id=P%5C_XGPgAACAAJ.

[51] Zhao Yanling, Deng Bimin, and Wang Zhanrong. "Analysis and study of perceptron to solve XOR problem". In: *The 2nd International Workshop on Autonomous Decentralized System, 2002.* 2002, pp. 168–173.

[52] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals, and Systems* 2.4 (Dec. 1989), pp. 303–314. DOI: 10.1007/bf02551274. URL: https://doi.org/10.1007/bf02551274.

[53] S. Sun et al. "A Survey of Optimization Methods From a Machine Learning Perspective". In: *IEEE Transactions on Cybernetics* (2019), pp. 1–14.

[54] Wei Hu et al. "Deep Convolutional Neural Networks for Hyperspectral Image Classification". In: *Journal of Sensors* 2015 (2015), pp. 1–12. DOI: 10.1155/2015/258619. URL: https://doi.org/10.1155/2015/258619.

[55] Matthew D. Zeiler and Rob Fergus. "Visualizing and Understanding Convolutional Networks". In: *Lecture Notes in Computer Science* (2014), pp. 818–833. ISSN: 1611-3349. DOI: 10.1007/978-3-319-10590-1_53. URL: http://dx.doi.org/10.1007/978-3-319-10590-1_53.

[56] Jost Tobias Springenberg et al. *Striving for Simplicity: The All Convolutional Net*. 2014. arXiv: 1412.6806 [cs.LG].

[57] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[58] Li Chen et al. "Adaptive Local Receptive Field Convolutional Neural Networks for Handwritten Chinese Character Recognition". In: *Pattern Recognition*. Ed. by Shutao Li, Chenglin Liu, and Yaonan Wang. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 455–463. ISBN: 978-3-662-45643-9.

[59] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Neural Information Processing Systems* 25 (Jan. 2012). DOI: 10.1145/3065386.

[60] Bing Xu et al. *Empirical Evaluation of Rectified Activations in Convolutional Network*. 2015. arXiv: 1505.00853 [cs.LG].

[61] Andrew L. Maas. "Rectifier Nonlinearities Improve Neural Network Acoustic Models". In: 2013.

[62] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Book in preparation for MIT Press. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[63] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. arXiv: 1412.6980 [cs.LG].

[64] George Bebis et al., eds. *Advances in Visual Computing*. Springer International Publishing, 2019. DOI: 10.1007/978-3-030-33723-0. URL: https://doi.org/10.1007/978-3-030-33723-0.

[65] D. Shen, G. Wu, and H. I. Suk. "Deep Learning in Medical Image Analysis". In: *Annu Rev Biomed Eng* 19 (June 2017). [DOI:10.1146/annurev-bioeng-071516-044442], pp. 221–248.