Quality Assurance of Image Registration in Radiotherapy using Deep Learning

Nasiq Ziyan

10119757

School of Physics and Astronomy

University of Manchester

Semester 2 MPhys Project Report

May 2022

This experiment was performed in collaboration with *Ahmed Maniyar (10401325)*

Abstract

Patients undergoing adaptive radiotherapy (ART) experience anatomical changes over the course of treatment. To minimise the radiation dose to organs at risk (OARs), clinicians manually delineate contours around these regions before each treatment session — a taxing clinical routine. However, these contours can be propagated throughout treatment using deformable image registration (DIR) but quality assurance is required for clinical implementation. A convolutional neural network was trained on 3D PET-CT data of the right parotid gland to predict the similarity between DIR-produced (registered) contours and manually drawn contours (ground truth), quantified by two similarity metrics known as DICE and the Hausdorff distance. The model's predictive performance was evaluated using the Pearson coefficient. The highest performance observed was when the model was trained to predict the DICE coefficient, producing a Pearson coefficient of 0.931 and 0.503 on the training and validation set, respectively. An occlusion map function was also developed to visualise the salient features learnt by the model, thereby providing a qualitative description of the model's predictive performance; however, numerous limitations prevented a justified interpretation of the map.

Contents

1 Introduction **3**

2 Project Background **4**
2.1 Image Registration 4
2.2 Application of Deformable Image Registration in ART 4
2.3 Similarity Metrics 5

3 Neural Networks in Deep Learning **7**
3.1 Dense Networks 7
3.2 Network Training 7
3.3 Convolutional Neural Networks 8
3.4 Overfitting and Network Performance Visualisation 10

4 Method **11**
4.1 Input Data 11
4.2 Model Architecture 12

5 Results and Discussion **12**
5.1 First Approach 12
5.2 Second Approach 16

6 Conclusion **19**

1 Introduction

Radiotherapy (RT) utilises ionising radiation, often in conjunction with chemotherapy and surgery, for effective cancer treatment [1]. X-rays are directed towards a tumour region using a medical linear accelerator (LINAC) that intends to cause tumour shrinkage by destroying cancerous cells (apoptosis) or disrupting cell division by damaging vital cell components, such as their DNA [2]. During treatment, surrounding healthy tissue is inevitably irradiated; however, DNA repair mechanisms can prevent mutations and thus, cancer cell proliferation [3]. Despite this, second malignancies are often observed due to unsuccessful DNA repair [4]. Therefore, treatment plans are designed to ensure that the dose conforms to the tumour region with minimal exposure to the surrounding tissue. Particular caution is directed towards organs at risk (OARs), such as the brain stem, whose irradiation may generate severe complications — especially as anatomical changes (internal motion, weight loss and ideally, tumour shrinkage) are frequently observed over the course of treatment [5].

Adaptive radiotherapy (ART) aims to counter this by refining a patient's treatment plan in response to their functional and anatomical changes [6, 7]. To achieve this, computer tomography (CT) scans are obtained during the pre-treatment stage where tumour volumes and OARs are delineated by a radiation oncologist as regions of interest (ROIs) [8]. However, different imaging modalities perceive tissue differently — in particular, magnetic resonance imaging (MRI) provides excellent soft-tissue visualisation while positron-emission tomography (PET) can present functional changes in the patient's body [4, 5, 9]. These can be leveraged for enhanced contrast in the ROIs. Images obtained from these modalities can be mapped to a single coordinate system using multi-modal image registration for refined volume delineation [9].

Image segmentation of ROIs is routinely performed manually or semi-automatically before each treatment session; as treatment periods may last up to 7 weeks, this procedure can be especially taxing on teams with limited clinical resources [8, 10]. Furthermore, the subjective nature of this task raises concerns regarding the susceptibility to intraobserver variability. Reports have associated the impact with lower overall survival rates [10].

However, recent computational advances provide a promising opportunity to optimise treatment planning through automating major components in the clinical workflow, such as contouring, dose calculation and quality assurance (QA) [9]. Automatic segmentation can be achieved using a deformable image registration (DIR) algorithm; with this technique, manual delineation is performed on a CT scan before the first session, and then deformed with respect to a CT scan obtained in the subsequent session, thereby simulating the anatomical changes [11]. This is known as contour propagation. However, QA is necessary to verify that automated contours using DIR are sufficiently similar to a radiation oncologist's manual segmentation, defined as the ground truth.

This project aims to accurately predict the similarity between a DIR-produced contour around the parotid gland (OAR) and its ground truth by constructing a convolutional neural network (CNN) and training it on 2D CT scans of parotid gland data. If the model is successful, predictions falling below a clinically set threshold would be indicative of poor contour propagation and thus, flagged to a clinical team for manual review. This project continues the work carried out last semester and by previous students [12, 13, 14, 15, 16].

2 Project Background

2.1 Image Registration

Image registration is a procedure of aligning identical anatomical points from two or more images [17]. This is achieved by transforming the first image (floating image) to match a second image (reference). The objective is to determine the optimal transformation that maximises the alignment of features in regions of interest. In RT, image registration may be first applied to images obtained from different modalities (multi-modal image registration) for improved volume delineation [9]. Then, this may be followed by obtaining images of a patient at different times (multi-temporal image registration) over the course of their treatment to simulate their anatomical variations [9]. This may necessitate modifications to the treatment plan for safer dose delivery.

Performing image registration requires two primary components: an objective function and a mathematical model [18]. The mathematical model describes the transformation between two images. Global transforms are typically classed as either rigid or affine while local transforms are referred to as non-rigid (deformable) [9]. Rigid transformations are parametrised by translations and rotations while affine transformations also include scaling and sheering. Non-linear transformations allow for local deformations by determining a displacement vector to each voxel x in the floating image and mapping it to the identical anatomical point in the reference. Thus, a displacement vector field $D(x)$ can be produced by compiling displacement vectors of all voxels in the floating image.

Deformable transformations are suitable for this project since expected anatomical variations, such as tumour shrinkage, are localised to particular regions. Thus, the deformable transformation of the floating image $I_F(x)$ to its reference $I_R(x)$ can be mathematically expressed as [11]

$$I_F(x + D(x)) = I_R(x), \tag{1}$$

The second component required for image registration is the objective function which represents the degree of alignment after applying one of the aforementioned transformations [19]. For example, a feature-based objective function measures the distance between the same anatomical landmark on each image. More commonly used is an intensity-based objective function which measures differences in voxel intensity patterns, based on the observation that common tissue types tend to have similar voxel intensities. Combining both components presents an optimisation problem of minimising the objective function which corresponds to an optimal registration [18]. This problem can be expressed mathematically as [17]: $\min_T [E(I_R, T(I_F)]$, where T is the transform and E is the objective function.

2.2 Application of Deformable Image Registration in ART

As the purpose of this project is QA of deformable image registration (DIR), several components are required as inputs for the CNN to output predictions. The acquisition procedure is summarised in this section [11, 20].

Firstly, a CT scan of a patient undergoing ART is obtained pre-treatment; this is defined as the floating image. Contours are manually drawn around OARs by a clinician denoting the regions where the radiation dose is to be minimised. In the subsequent treatment session, before dose delivery, a second scan is obtained which is defined as the reference. Registering both images using DIR produces a DVF that maps the anatomical variations between the two

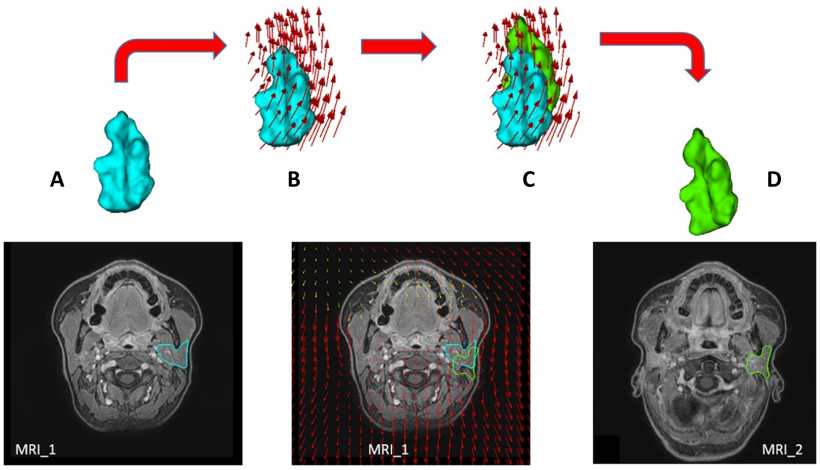

Figure 1: An illustration of applying contour propagation to the right parotid gland on an MRI image. The floating and reference image are labelled MRI_1 and MRI_2, respectively [21].

sessions. Applying the DVF to the floating image contour deforms it in accordance with the anatomical variations, referred to as the registered contour. In this project, contour propagation will be applied to the parotid gland as shown in Figure 1 [21].

QA of DIR entails a comparison between the registered contour and an ideal 'ground truth' contour. The latter is obtained by manual delineation of the OARs on the reference image. To quantify the similarity between the registered and ground truth contours, an appropriate similarity metric must be chosen. While successful clinical implementation of this procedure will accelerate workflow, the limitations of current DIR algorithms may outweigh these benefits. For example, erroneous registration may occur when registering two images where an anatomical feature exists on one image but not the other [9]. Furthermore, image registration outputs are limited in resolution due to the finite number of voxels. However, improving resolution may not be effective overall as an overly complex dataset may be unfeasible for deep learning applications [9, 22, 23].

The reference image, registered contour, DVF and similarity metric data are the required components for this project's deep learning application. However, the choice of similarity metric needs careful consideration based on the clinical context.

2.3 Similarity Metrics

Numerous similarity metrics are available in the medical field, typically categorised into six groups [24, 25]: probabilistic based, volume based, information-theoretic based, pair-counting based, overlap based and distance based. The relevant types for this project are overlap based and distance based. Firstly, the DICE coefficient is a symmetric, overlap based metric and is most commonly used for measuring the similarity between two volumes in image segmentation applications, defined as [21, 26]

$$\mathrm{DICE}\,(A, B) = \frac{2\,(A \cap B)}{|A| + |B|} \tag{2}$$

where A and B are the two volumes. The metric is symmetric under the exchange of these volumes and normalised between 0 and 1 where 1 indicates a complete overlap. In this project, A and B represent the voxels enclosed within the manually drawn and registered contour, re-

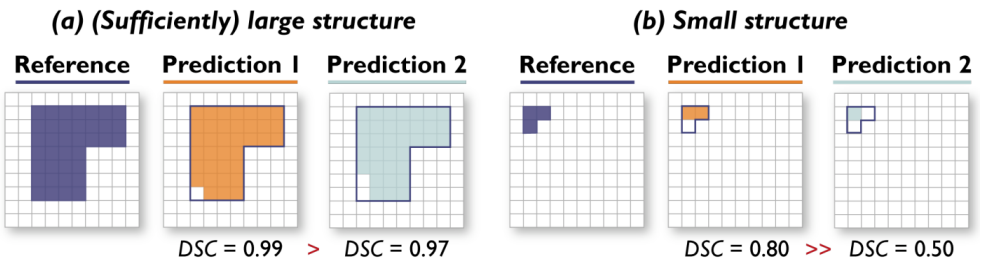

Figure 2: Impact of object size on the DICE coefficient. In both scenarios, the 1st and 2nd prediction differ by 1 pixel, but the impact on the DICE coefficient varies significantly [27].

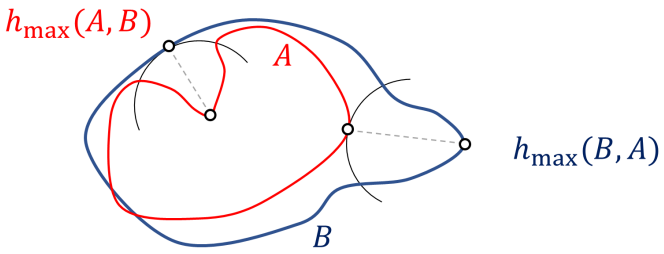

Figure 3: A diagram of the directed Hausdorff distances from A to B ($h_{\max}(A,B)$) and vice versa ($h_{\max}(B,A)$). The directed HD is not necessarily symmetric as $h_{\max}(A,B) \neq h_{\max}(B,A)$ [12].

spectively. Since the DICE coefficient is insensitive to the absolute volume, problems may be encountered in this clinical application. DICE may be particularly unsuitable in applications involving the delineation of small regions relative to voxel size, such as a brain lesion, as illustrated in Figure 2 [27].

The DICE coefficient also makes no consideration for contour perimeters. This may be a crucial flaw in RT applications as OARs may be complex in structure [27]. Therefore, radiation exposure to an unenclosed region of an OAR may produce detrimental outcomes. A distance based metric such as the Hausdorff distance (HD) may address this limitation [28]. The directed HD from a contour A to a second contour B is defined as the maximum of all distances from a point on contour A to the nearest point on contour B [29]. This is illustrated in Figure 3 and expressed as [25]

$$h_{\max}(A,B) = \max_{a \in A} \min_{b \in B} || \, a - b \, ||, \tag{3}$$

where $|| \, a - b \, ||$ is the distance between a point on contour A and a point on contour B. This can be made symmetric by taking the maximum of the two directed HDs:

$$\mathrm{HD}(A,B) = \max(h_{\max}(B,A),\, h_{\max}(A,B)). \tag{4}$$

Noise and artefacts are commonly observed on medical data which can be interpreted as outliers during image segmentation and lead to registration errors. This can be solved by replacing the maximum distance in Equations 3 and 4 with the mean, defined as mean directed HD:

$$h_{\text{mean}}(A, B) = \frac{1}{N} \sum_{a \in A} \min_{b \in B} || \, a - b \, ||. \tag{5}$$

Alternatively, the 95th percentile of all distances can be computed, referred to as HD95, and is the most commonly used variant [27]. For all HD metrics, a value of 0 refers to perfect alignment and no fixed upper bound exists; however, a normalisation from 0 to 1 is typically applied [27].

3 Neural Networks in Deep Learning

Deep learning (DL) provides a solution to automated feature learning [30]. In this project, we aim to construct a model that infers the similarity between a registered and ground truth contour. It achieves feature learning by correlating a reference image, a corresponding DVF and a registered contour (inputs), with the true similarity between the two contours (label). The procedure is summarised in this section.

3.1 Dense Networks

A dense neural network (NN) has the simplest configuration for a DL network which eases the understanding of the essential network components [31]. It is comprised of an input layer followed by a series of hidden layers that output to a final layer [32, 33]. Neural networks whose goal is to predict a discrete class will have a number of nodes in the output layer equal to the number of classes. Whereas, a network tasked with predicting one or more continuous values will have a node for each continuous value. The latter is referred to as a regression network which is the relevant type for this project.

Each hidden layer may contain several nodes that are tasked with extracting increasingly complex features from the preceding layer. For example, a given layer may detect corners and shapes from having learnt single edges in a preceding layer. If all nodes in a given layer are connected to all nodes in the layers before and after it, the network is defined as dense.

All connections between nodes have an associated weight which represents the connection strength. The output of a node in a prior layer multiplied by the weight connection to a node in the proceeding layer serves as one of the inputs to the proceeding layer's node. The total input is computed by summing over all nodes in the prior layer. This input is passed to an activation function that converts it into an output. The larger the activation of a node, the greater impact it will have on the final prediction. A full mathematical description of this process is outlined in the first semester report [12].

There are numerous activation functions available; however, the appropriate choice is dependent on the function which maps the input data to the output labels. Nonlinear activation functions are typically employed as this is the expected relationship in most deep learning applications [32]. In this project, the nonlinear activation function chosen is the rectified linear unit (ReLU), shown in Figure 4 [12].

3.2 Network Training

Training a neural network first involves passing input information throughout the network to the final layer for prediction (forward propagation) [32, 34]. Input data is often fed in batches, typically 32, to improve computational efficiency. One pass of all batches comprising the entire data input is defined as an epoch. For a single sample in a batch, a comparison is made between the prediction and the label using a loss function C; this produces a loss value, where a high loss

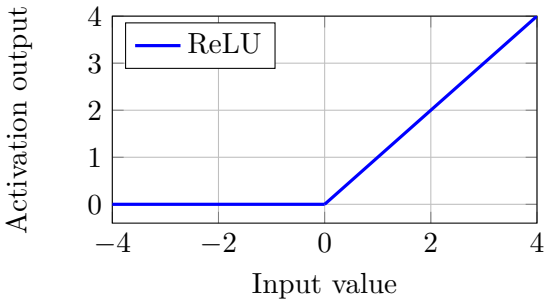

Figure 4: A graph of the nonlinear rectified linear unit (ReLU) activation function [12].

implies a poor prediction. The most common loss function is the mean squared error (MSE), defined as [32]

$$C_{\text{MSE}} = \frac{1}{N}\sum_{i=0}^{N}(y - y_i)^2, \tag{6}$$

where y is the label, y_i is the model prediction and N is the number of samples in the batch. Then, the model updates the network weights aiming to improve predictions for the next batch. Two algorithms are required to achieve this [32]. Firstly, the backpropagation algorithm calculates the gradient of the loss with respect to the network's trainable parameters [35]. The network learns by minimising the loss using a gradient descent optimisation algorithm, such as stochastic gradient descent (SGD) [32]. By representing all the old network weights as a weight vector $\mathbf{w}$, the updated weights $\mathbf{w}'$ are determined using [32]

$$\mathbf{w}' = \mathbf{w} - \eta\nabla_{\mathbf{w}}C, \tag{7}$$

where η is the learning rate, a hyperparameter that dictates the model's step size taken in the direction of decreasing loss. An explanation of the intuition behind backpropagation using a mathematical representation can be found in the first semester report [12].

3.3 Convolutional Neural Networks

Another branch of neural networks, designed especially for computer vision tasks, is convolutional neural networks (CNNs) [36]. As with dense networks, each hidden layer outputs to the layer ahead of it. Therefore, CNNs display a hierarchical structure in feature complexity, where earlier layers are generally capable of edge detection while later layers can use these to construct complex structures; however, unlike dense networks, CNNs are not fully connected. The primary distinctions of a CNN are the addition of convolutional and pooling layers [37].

Convolutional layers use filters (kernels), analogous to the nodes in a dense layer, for feature extraction via a convolution operation [33]. This results in a feature map output that highlights the extracted feature [37]. These layers are typically followed by an activation layer to introduce nonlinearity into the learning procedure. For a 2D CNN, a 2D filter (typically 3 x 3) containing a matrix of weights convolves a larger image by taking a series of horizontal and vertical steps (strides) until the entire image has been convolved.

While these weights can be chosen by the user, they are often randomly initialised to have a zero mean and a standard deviation of one [32]. These weights are updated iteratively via SGD and backpropagation. The filter size f and stride s_c can also be tuned by the user depending on

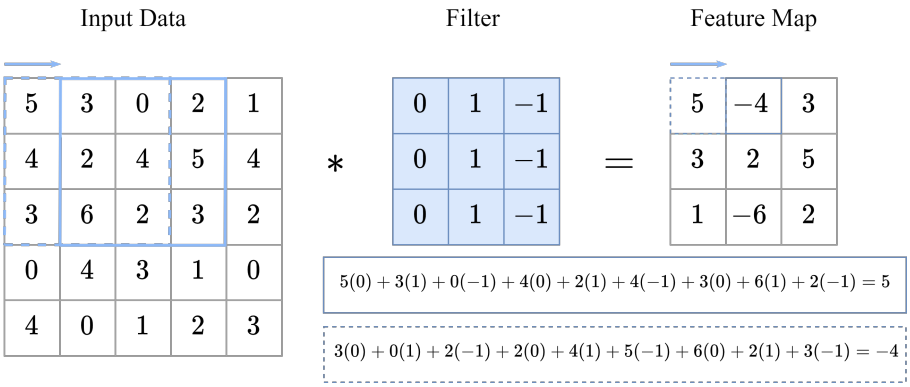

Figure 5: Visualisation of the convolution operation with a filter of size 3 x 3 convolving a 5 x 5 image with stride $s_c = 1$. Calculations for the first two elements in the feature map are provided (summation of the element-wise products) [12].

feature size and quality relative to the size of the input image. Each value in the feature map matrix is computed by a sum of the element-wise products between the filter and the region enclosed by the filter [38]. Figure 5 shows a visualisation of this operation [12].

As observed in this figure, the feature map is not the same size as the original input. Therefore, the size of feature maps in later convolutional layers may be significantly smaller, which may hinder effective feature extraction. While smaller kernels can prevent rapid shrinking with each convolutional layer, zero padding enables the preservation of the original image size [32]. It accomplishes this by adding a border of zero value elements around the edges of the image which increases the number of horizontal and vertical steps of the filter. From an input image with size M x M, the output feature map size M' x M' after convolution can be determined by

$$M' = \frac{M - f + 2P}{s_c} + 1, \tag{8}$$

where P is the padding (thickness of zero value border).

Input images are passed to the network as a tensor object, typically containing numerous channels per pixel, such as three colour channels for RGB, or for this project, two channels for the x- and y-components of the DVF. A user may employ multiple filters to convolve each channel separately, capable of performing channel-specific feature extraction, at the cost of increased computation [39]. To counter this, a subsampling layer, referred to as pooling, follows the activation layer, which attempts to summarise the features in the feature map [40].

Since feature extraction may be heavily dependent on the feature's location in the image, downsampling can extract the salient features while discarding the irrelevant background features and redundancies. Thus, pooling layers provide local translational invariance which ultimately makes the model less sensitive to image samples modified by a partial rotation, shift or any other minor augmentation [32, 40]. The pooling operation is applied in a similar approach to convolution. A filter of size p x p moves horizontally and vertically with stride s_p across a feature map, and an operation is performed on the enclosed values. The most common operation is computing the max value, referred to as max pooling, as shown in Figure 6. However, average pooling may be more appropriate for certain DL applications [40].

Generally, the model's weights are randomly initialised and normally distributed [40]. Particular initial configurations may consist of large outlying weights that may cause excessively large

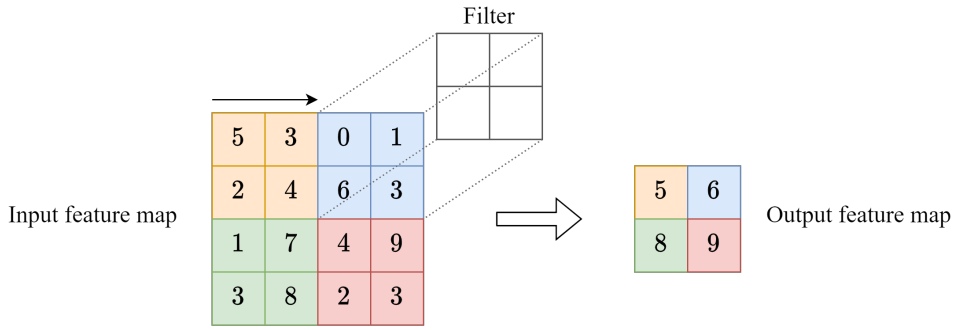

Figure 6: Illustration of max pooling applied to a 4 x 4 image using a filter with size $p = 2$ and stride $s_p = 2$. Adapted from [12].

activation outputs which can cascade throughout the network, resulting in network instability. This can be minimised using a batch normalisation layer which normalises the input weights on a per-batch basis, improving the efficiency and stability of the learning process.

3.4 Overfitting and Network Performance Visualisation

A network is first trained on a training set, reporting a loss value for each batch. This is then averaged over all batches in a single epoch to produce the final training loss for that epoch [32]. To continuously measure the network's predictive performance on unseen data, a validation set can be provided simultaneously to the model. This outputs a validation loss which is calculated at the end of each epoch as opposed to during it.

The validation set also provides a tool for monitoring overtraining. Ideally, we aim for similar training and validation loss that decreases with the number of epochs; however, the validation loss frequently increases during the later stages of training. This is referred to as overfitting and is indicative of poor generalisability [32]. To reduce overfitting, it is often useful to visualise the features learnt by a model. Numerous recent studies in the medical domain have demonstrated substantial advances by using occlusion-based mapping for CNN performance visualisation and interpretation [41]. The construction procedure of an occlusion map is summarised below.

Firstly, a portion of an input image is blocked out using a blocking filter, and a prediction value is computed for the entire image. The initial portion begins in the top left corner of the image. Then, the blocking filter is shifted right and a prediction value is computed on this particular image. Like this, the filter is subsequently shifted horizontally, computing a prediction at each step, until every portion of the image has been individually blocked out. Thus, a prediction value for each image with a different occluded region is computed and can be organised into a matrix. The spatial position of each element in the matrix corresponds to the region occluded before prediction. Subtracting the prediction value with no applied occlusion (reference prediction) from all elements in the matrix and plotting the final matrix results in the final occlusion map.

Figure 7 illustrates an example for a classification network using data obtained from the MNIST database [42]. As observed in Figure 7(c), the prediction probability for the correct class falls significantly (from red to blue) when regions in the centre are individually occluded. This is because the correct digit class under some central occlusion begin to imitate other incorrect digits. Since no signs of overfitting were observed during training, this occlusion map is indicative of the desired behaviour.

If overfitting is observed, corrective measures can be suggested based on the occlusion map,

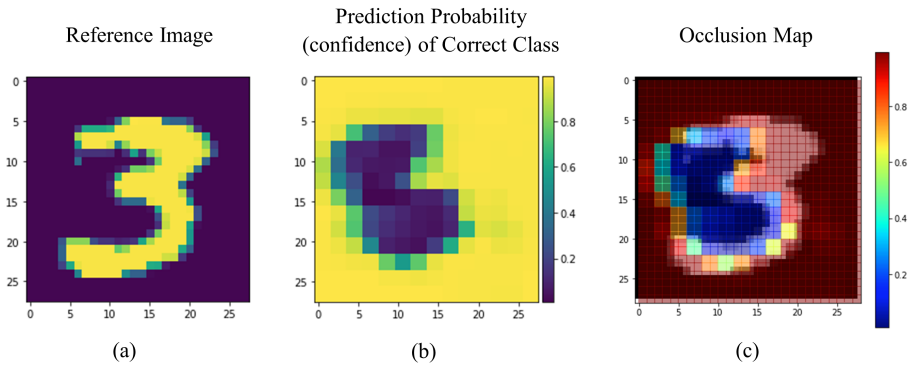

Figure 7: Illustration of an occlusion map example for a classification network. (a) Reference image sample obtained from the MNIST dataset [42]. (b) Prediction probability of correct class. 10 classes exist, corresponding to handwritten digits of values ranging from 0 to 9. (c) Occlusion map constructed by superposing (a) and (b).

such as data augmentation [43]. This may involve applying a perturbation to the input image, such as a small rotation, translation or blurring, in the regions where the prediction probability falls unexpectedly. The perturbed images are fed to the network with the same label, thereby simulating the acquisition of more data, and allowing the model to improve generalisability in the region of interest. This is further elaborated in Section 5.

4 Method

4.1 Input Data

The 3D PET-CT data for training and validation was obtained from the Head-Neck Cetuximab collection from the Cancer Imaging Archives [44]. The 3D data was sliced axially to produce 2D scans, each with pixel dimensions 512 x 512. On each scan, contours of the left and right parotid gland, brainstem and spinal cord were manually drawn by a clinician. However, only the right parotid gland was investigated in this semester of the project. As each slice contained a large irrelevant background, pixel dimensions were trimmed down to 256 x 256, reducing computational load.

As mentioned in Section 2.2, DIR in ART requires two scans of a single patient taken at different times, which produces a DVF that describes the anatomical change over this time interval. However, due to the limited amount of data, this procedure was simulated by pairing different patients together, where one patient's data is defined as a floating image and the other as a reference. 2162 pairs were registered which corresponds to 23679 registered slices. Of these 23679 slices, 14353 (60.6%) and 9326 (39.4%) were allocated to the training and validation set, respectively. Each slice has an associated DVF, a registered contour (produced by applying the DVF on the floating image's contour) and multiple similarity metrics (measured between the reference and registered contour). The two similarity metrics investigated in this semester were DICE and HD95.

Two approaches with regard to the model's input were investigated this semester. In the first approach, each pixel in the input data consisted of one brightness channel, corresponding to the reference image, and two DVF channels (x- and y-components). Therefore, no explicit contours were provided to the model. Anticipating that the model would experience difficulty in

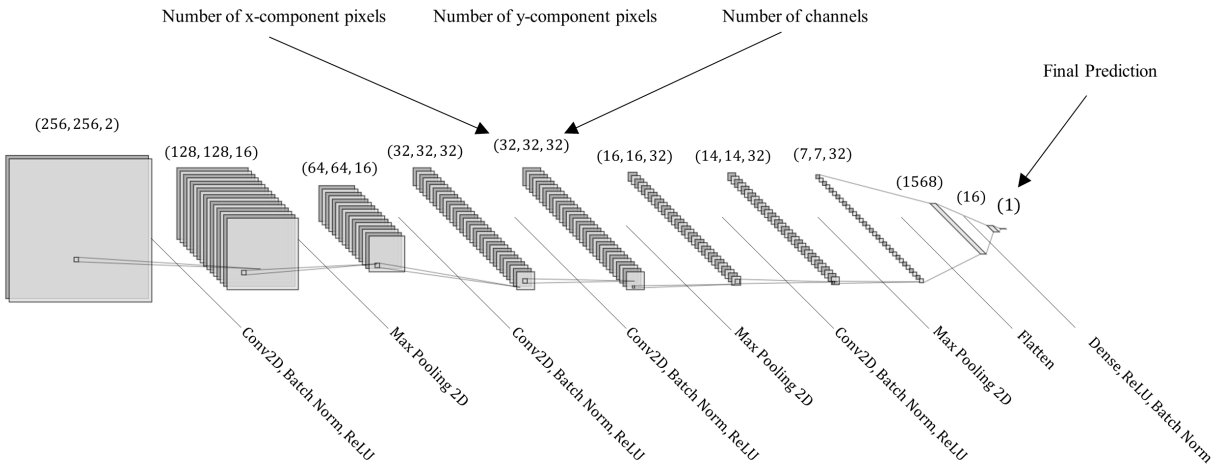

Figure 8: An illustration of the model's architecture used in this semester, created with a visualisation tool found in [45]. Layer types and feature map dimensions are illustrated on the diagram.

feature learning from the DVF due to significant noise and complexity in irrelevant regions, the second approach would only provide two brightness channels in total: the first corresponds to the reference image while the second corresponds to the registered contour, where a brightness value of 1 defines the contour perimeter and a value of 0 is assigned everywhere else. Thus, it was hoped the model would be assisted in feature localisation by the superposition of the registered contour.

In the first semester, a range of data pipeline and memory management techniques, such as memory maps and 'Datasets', were introduced due to VRAM limitations [12, 13]. However, besides the memory maps, these changes were reverted since the trimming of pixel dimensions reduced the data's file size to a feasible value.

4.2 Model Architecture

The model's architecture is illustrated in Figure 8. All convolutional layers had filter size 3 x 3 with stride $s_c = 2$, and all max-pooling layers had size 2 x 2 with stride $s_p = 2$. The batch size was 32 and the optimiser choice was 'Adam' (adaptive moment estimation) with learning rate $\eta = 0.001$. The loss function employed was the mean squared error (Equation 6). It was anticipated that regions with the highest brightness in the reference image had an excessive influence on the final prediction, leading to poor validation accuracy. Therefore, we took the natural logarithm of all brightness values, followed by a normalisation of [0,1].

5 Results and Discussion

5.1 First Approach

In our first approach, the input data was a tensor of shape (23679, 256, 256, 3), corresponding to the number of 2D slices, pixels in the x- and y-axis, and channels, respectively. The 3 channels consisted of the reference image's brightness and the x- and y-components of the DVF. As a zero DICE score corresponds to a zero overlap between the registered and ground truth contour around

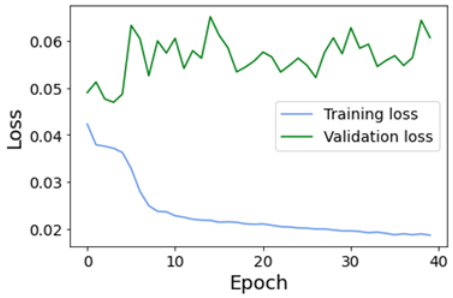

Figure 9: Plot of the DICE training and validation loss calculated for 40 epochs. No early stopping was used, therefore, the minimum potential training loss was likely not reached.

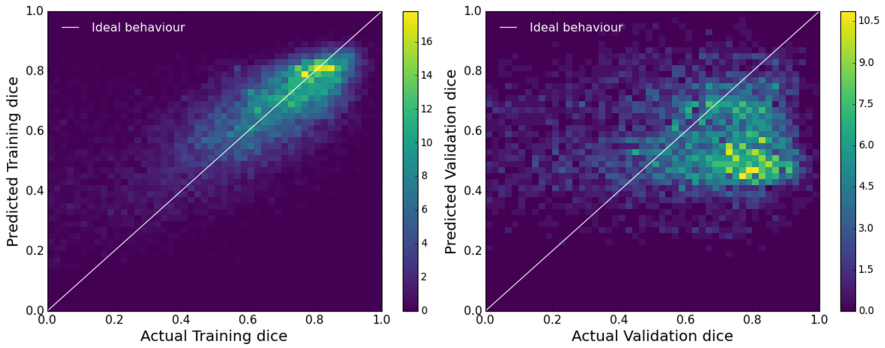

Figure 10: Heatmaps illustrating the correlation between the model's predicted DICE against the actual DICE (ground truth) for the training (left) and validation set (right) using the first approach. The Pearson coefficients for the training and validation set were 0.715 and 0.05, respectively.

the right parotid gland, it was anticipated that the model would struggle to learn a correlation between the input data and labels, given that no contours were present in the input data. That is, if two or more highly irregular DVFs (but different from one other) had a DICE similarity of zero, the model would likely learn irrelevant correlations between the DVFs. Therefore, all zero DICE scores were removed before training and validation. The training and validation loss as a function of epoch is shown in Figure 9. Subsequently, heatmaps comparing the model's predicted DICE metric against the actual DICE were plotted for the training and validation set, shown in Figure 10.

Ideally, we expect to observe a linear relationship between the predicted and actual DICE similarity. The linearity between two data sets can be measured by the Pearson coefficient, ranging from -1 to 1, which describes the strength and direction of the correlation but is independent of the gradient [46]. A value of 1 indicates a perfect positive correlation, a value of -1 indicates a perfect negative correlation and 0 indicates no linear correlation. The Pearson coefficients for the training and validation set were determined as 0.715 and 0.046, respectively. The large discrepancy between the two coefficients indicates overfitting to the training set and a lack of generalisability which is substantiated by the discrepancy between the two loss curves in Figure 9. In fact, the validation loss curve slightly increased with the decreasing training loss, suggesting that the features learnt were firmly restricted to the training set.

The saliency of learnt features in the input image can be visualised by generating an occlusion

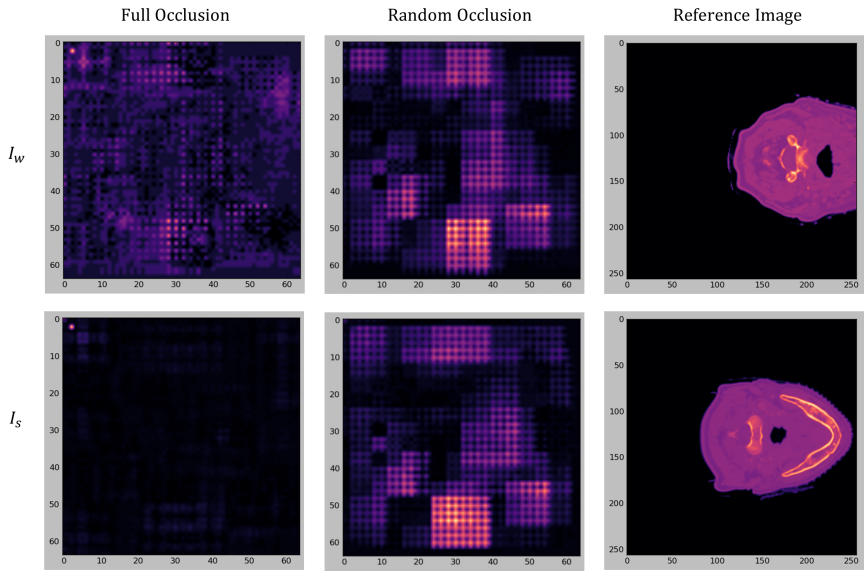

Figure 11: Occlusion maps for Images I_w (top) and I_s (bottom), corresponding to weak and strong DICE predictions, respectively. For image I_w: $v_{p,w} = 0.736$ and $v_{a,w} = 0.244$, thus, $d_{max} = 0.492$. For I_s: $v_{p,s} = 0.311$ and $v_{a,s} = 0.239$, thus, $d_{min} = 0.072$. Domains D_1 and D_2 were set as [0.2, 0.3] and [0.234, 0.254], respectively.

map for images within the validation set. The blocking filter size for all occlusion maps was chosen to be 2 x 2. Two types of perturbations within the blocking filter were investigated. In the first type, the channel values for all pixels within the blocking filter were set to zero, termed full occlusion (or 'zero' perturbation). In the second, the values for the reference image brightness were randomised from [0, 1] whereas the values for the DVF channels were randomised from [-1, 1], termed random occlusion. Bright regions on an occlusion map indicate a large absolute difference between predictions in the occluded and unoccluded image.

In order to compare occlusion maps between strong and poor predictions, the following approach was used. Firstly, a domain D_1 on the actual validation dice values was chosen, such as [0.4, 0.6]. Then, the difference between the predicted validation $v_{p,w}$ and actual validation $v_{a,w}$ for all images within this domain was computed. The image with the largest absolute difference (d_{max}) was defined as the weak prediction, I_w. A narrower domain D_2 was then centred at $v_{a,w}$. Within this narrow domain, the image with the minimum absolute difference (d_{min}) between predicted validation $v_{p,s}$ and actual validation $v_{a,s}$, was defined as the strong prediction image, I_s. Thus, the two images for comparison will have similar v_a values but significantly different v_p values. Figure 11 shows the occlusion maps for images I_w and I_s.

On all occlusion maps, a grid-like structure of varying sizes was observed. This likely corresponds to the influence of filters whose size ranges from 128 x 128 to 7 x 7, as shown in Figure 8. The varying brightness of these structures likely corresponds to the influence of the varying activation strengths of these filters. On Image I_w, large regions with similar intensity were observed for both perturbation types which may indicate the model's inability to distinguish the salient regions from the irrelevant regions in the reference image. This substantiates the d_{max} value of 0.492. For I_s, a similar occlusion map under random perturbations was observed; however, the full occlusion map indicates insensitivity to 'zero' perturbation. While it is unclear why this has occurred, a visualisation of the DVF may provide insight. For example, if the DVF components

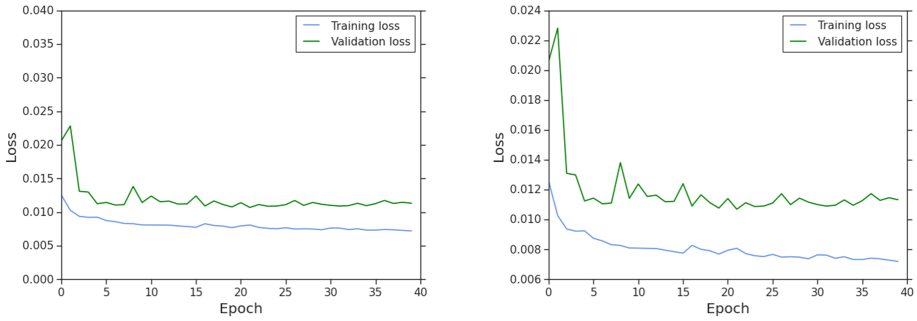

Figure 12: Left: Plot of the HD95 training and validation loss calculated for 40 epochs. Right: a magnified version of the left image.

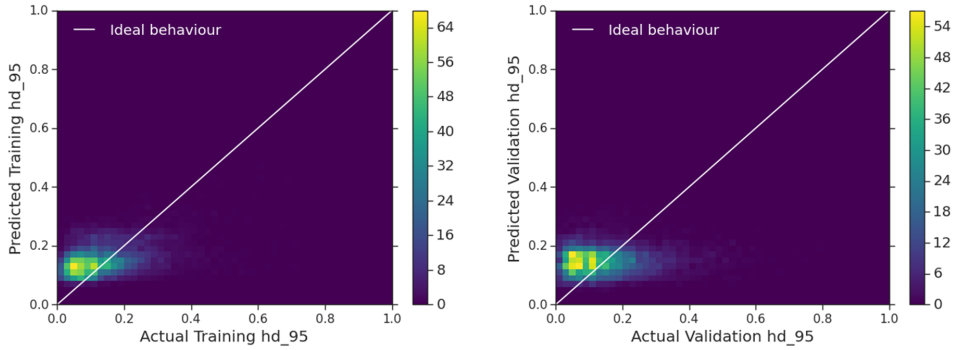

Figure 13: Heatmaps illustrating the correlation between the model's predicted HD95 against the actual HD95 (ground truth) for the training (left) and validation set (right) using the first approach. The Pearson coefficients for the training and validation set were 0.419 and 0.0496, respectively.

were relatively small or zero in most regions, a 'zero' perturbation will have a minimal effect on the DVF, and thus the prediction differences plotted in the occlusion map will be minimal. Despite the low d_{min} of 0.072, the low Pearson validation of 0.05 indicates that this strong prediction was due to chance as opposed to salient feature learning. Therefore, interpreting occlusion maps is likely inappropriate for poor Pearson validation cases.

The same approach was applied to HD95. Figure 12 and 13 shows the loss curves and heatmaps, respectively.

Despite HD95 achieving a lower training and validation loss, the Pearson training and validation were both lower than DICE at 0.419 and 0.045, respectively. Figure 14 shows the occlusion maps for HD95. Unlike DICE, faint delineations between the background and foreground were observed, especially under full occlusion, suggesting partial recognition of the anatomical structure.

A significant limitation of this occlusion map method is the lack of constraint on the determined axial slice; images I_s and I_w are highly dissimilar which suggests that an occlusion map comparison is unsuitable. The moderate Pearson training scores paired with the weak Pearson validations for both metrics in the first approach suggest that the DVF's complexity hinders generalisability and localisation of salient features. Our second approach was an attempt to rectify this.

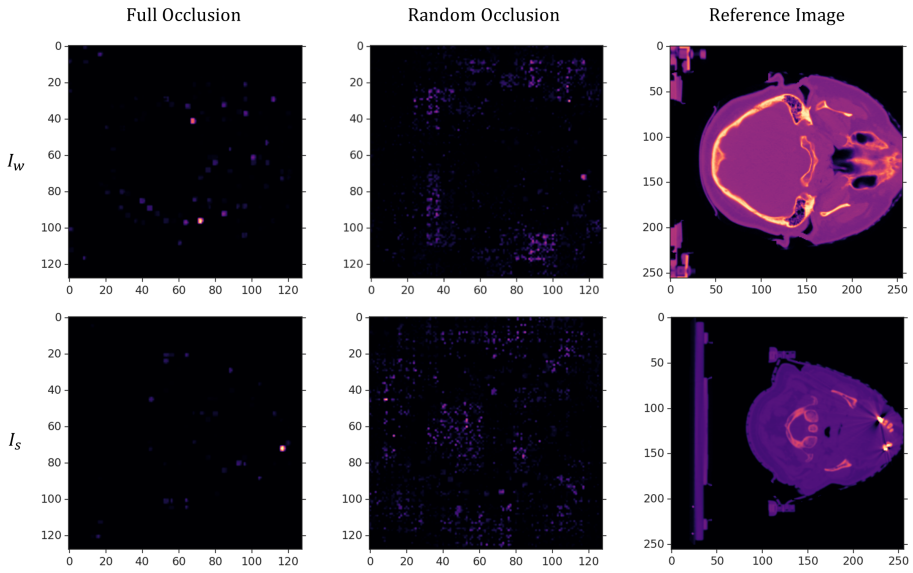

Figure 14: Occlusion maps for Images I_w (top) and I_s (bottom), corresponding to weak and strong HD95 predictions, respectively. For image I_w: $v_{p,w} = 0.350$ and $v_{a,w} = 0.057$, thus, $d_{max} = 0.293$. For I_s: $v_{p,s} = 0.0798$ and $v_{a,s} = 0.0570$, thus, $d_{min} = 0.0227$. Domains D_1 and D_2 were set as [0.04, 0.07] and [0.055, 0.059], respectively.

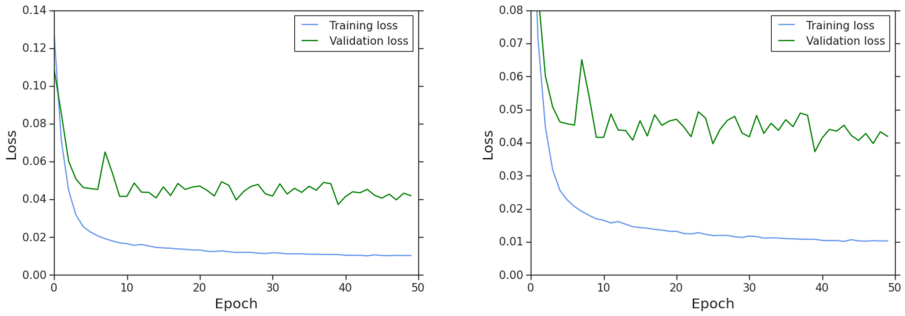

Figure 15: Training and validation loss curves for DICE over 40 epochs.

5.2 Second Approach

The only change was to the input data in which the 2 DVF channels were replaced by a binary channel denoting the registered contour. Figure 15 and 16 shows the loss curves and heatmaps, respectively.

Both the training and validation loss stabilised to a lower value compared to the first approach. As a result, strong Pearson training and validation scores of 0.931 and 0.503 were achieved, respectively. This indicates that the removal of the DVF sufficiently reduced complexity such that the features learnt in the training set were relatively generalised. Figure 17 shows the occlusion maps for DICE.

Despite Image I_w having a large d_{max} of 0.647, the model successfully delineated the background from the foreground which may suggest that the model's feature maps are minimally influenced by full and random perturbations in the background. Little to no brightness was observed within and around the registered contour; this is undesirable for a perfectly trained

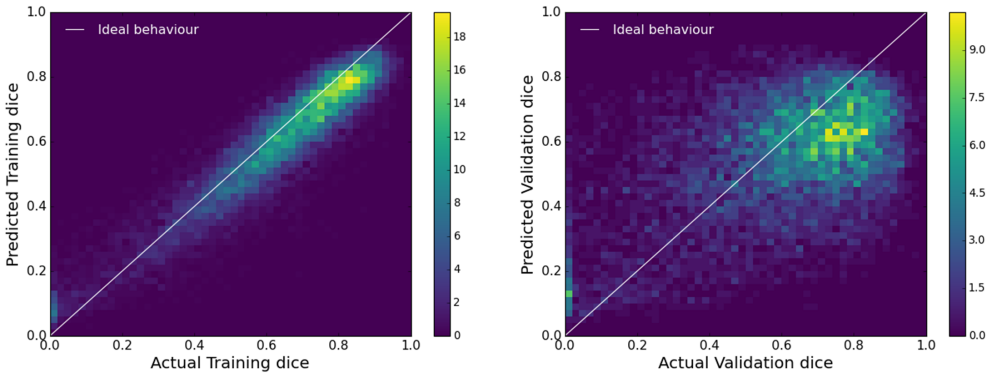

Figure 16: Heatmaps of the correlation between the model's predicted DICE against the actual DICE (ground truth) for the training (left) and validation set (right) using the second approach. The Pearson coefficients for the training and validation set were 0.931 and 0.503, respectively.

model but expected for a 0.503 Pearson validation score. Since no axial slice constraint on I_s given I_w was imposed, the registered contours are significantly different. Additionally, the DICE metric is unsuitable for relatively small contour areas, as explained in Section 2.3. Therefore, interpretations between I_s and I_w, and I_s itself, are inappropriate.

Although not presented, the occlusion maps for I_s all had relatively small contours for most D_1 domains. Since relevant data provided to the model is negligible in these cases, our findings suggest that there is a high probability of finding a strong prediction ($d_{min} \lesssim 0.1$) based purely on chance, for any D_1 with domain width ~ 0.2. This is substantiated by I_s under random occlusion where a lack of localisation and background sensitivity was observed (Figure 17). Imposing a constraint on the minimum contour area when determining I_s and I_w may increase the benefit of occlusion map interpretation.

Figures 18 and 19 show the loss curves and heatmaps for HD95, respectively. The Pearson coefficients for the training and validation set were 0.706 and 0.273, respectively, indicating relatively good model performance. The training loss stabilised at approximately 0.06 (0.08 in the first approach) whilst the validation loss fluctuated around 0.011 (0.012 in the first approach). Therefore, the second approach proved substantial performance improvements for both metrics but more so for DICE. However, the occlusion maps for HD95, shown in Figure 20, display the desired occlusion map.

Under both occlusion types, Images I_w and I_s demonstrated localisation of salient features. It was observed that, within the contours of I_s, the number of pixels and their brightness was generally higher than I_w. Additionally, the opposite behaviour was observed outside of the contours. Both of these observations are the expected result for a well-trained model. While all investigated domains displayed similar results, a quantitative approach is necessary to confirm these observations.

A key limitation of the occlusion maps presented is with regards to the calculation of each pixel's brightness in the map; as explained in Section 3.4, the brightness of each pixel is determined by the absolute difference in prediction between the occluded and unoccluded image. Therefore, no distinction was made on whether the prediction moved towards or away from the actual value (ground truth). If implemented, this distinction can be visualised on an occlusion map by categorising pixels into two groups based on the direction of deviation and weighting them by the extent of deviation. Combining this with the aforementioned suggestions, such as the axial slice constraint, will allow for better justification of interpretation.

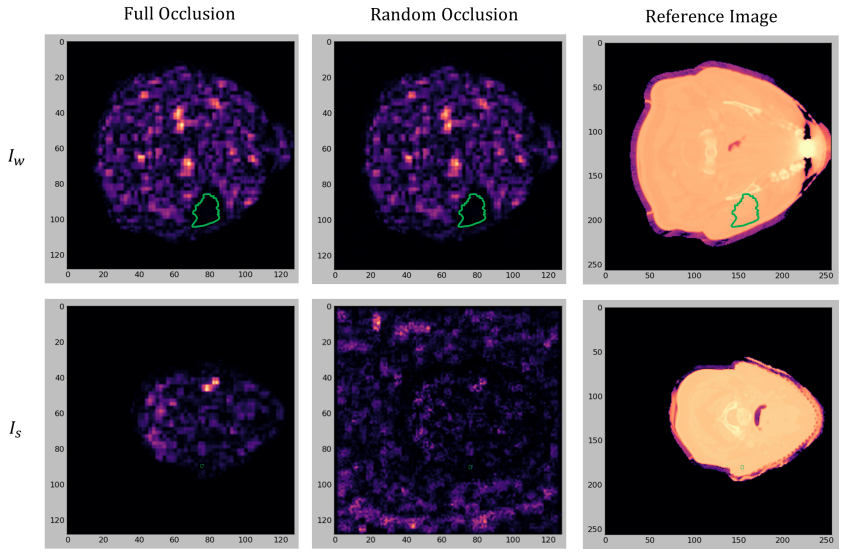

Figure 17: Occlusion maps for Images I_w (top) and I_s (bottom), corresponding to weak and strong DICE predictions, respectively. For image I_w: $v_{p,w} = 0.725$ and $v_{a,w} = 0.078$, thus, $d_{max} = 0.647$. For I_s: $v_{p,s} = 0.127$ and $v_{a,s} = 0.0848$, thus, $d_{min} = 0.0422$. Domains D_1 and D_2 were set as [0.06, 0.08] and [0.068, 0.0088], respectively. Registered contours are annotated in green.

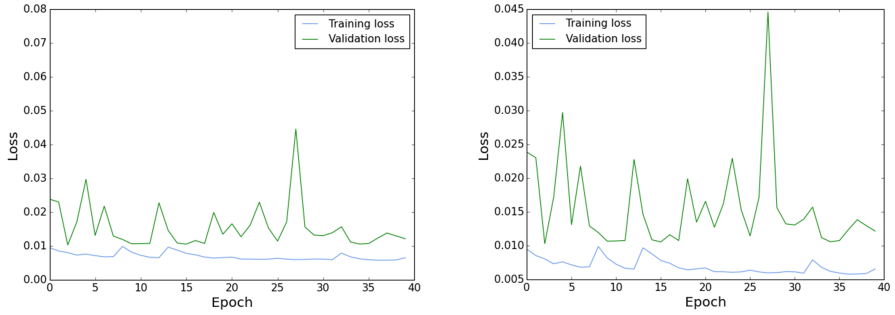

Figure 18: Left: Plot of the HD95 training and validation loss calculated for 40 epochs. Right: a magnified version of the left image.

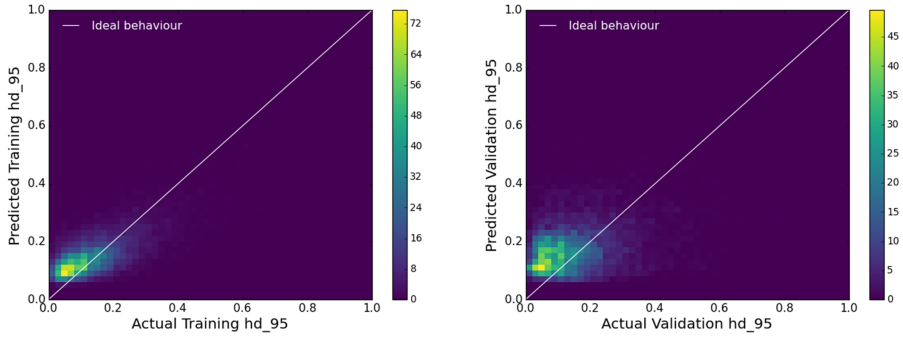

Figure 19: Heatmaps illustrating the correlation between the model's predicted HD95 against the actual HD95 (ground truth) for the training (left) and validation set (right) using the second approach. The Pearson coefficients for training and validation were 0.706 and 0.273, respectively.

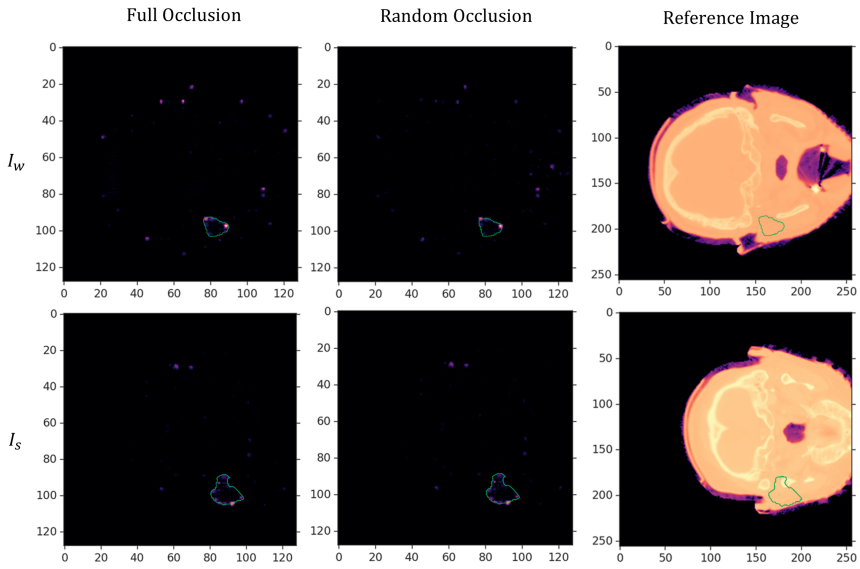

Figure 20: Occlusion maps for Images I_w (top) and I_s (bottom), corresponding to weak and strong DICE predictions, respectively. For image I_w: $v_{p,w} = 0.383$ and $v_{a,w} = 0.050$, thus, $d_{max} = 0.333$. For I_s: $v_{p,s} = 0.0728$ and $v_{a,s} = 0.050$, thus, $d_{min} = 0.0228$. Domains D_1 and D_2 were set as [0.04, 0.07] and [0.0475, 0.0515], respectively. Registered contours are annotated in green.

6 Conclusion

In this project, QA of a deformable image registration (DIR) algorithm was attempted using a 2D convolutional neural network. This network was trained on two different inputs provided by DIR to predict two similarity metrics, DICE and HD95. The prediction performance was quantified by the Pearson coefficient where a strong positive correlation for training and validation indicates good performance. With the first input type, the model received 23679 slices with three channels corresponding to the brightness of a reference image and the x- and y-components of an associated DVF. The Pearson coefficients on the training and validation sets for DICE were 0.715 and 0.046, respectively. For HD95, both scores were lower at 0.419 and 0.045. The poor validation scores are indicative of poor generalisability, likely caused by the significant complexity of the DVF. An occlusion function was developed to visualise the impact on the final prediction when small perturbations are applied to the input data. Bright regions on the produced occlusion map can be interpreted as salient features from the model's perspective. However, numerous limitations of the occlusion function indicated its unsuitability for a qualitative diagnosis of the model's predictive performance, especially in low Pearson validation cases.

With the second input type, the DVF component channels were replaced by a binary brightness channel corresponding to the registered contour. Significant improvements in prediction performance for both metrics were observed but more so for DICE. The Pearson coefficients on the training and validation sets for DICE were 0.931 and 0.503, respectively. Meanwhile, for HD95, the coefficients computed were 0.706 and 0.273, respectively. While HD95 was lower than DICE for both coefficients in the second approach, occlusion maps for HD95 displayed the desired behaviour to a greater extent than DICE. However, since only a tiny proportion of scans were investigated, this observation is inconclusive of the holistic behaviour. Thus, a quantitative occlusion metric which can be computed for all scans is necessary for a rigorous justification.

Overall, the removal of the DVF in favour of the registered contour and predicting the DICE metric produced the best training and validation performance. However, the associated Pearson coefficients still fall short for clinical implementation in adaptive radiotherapy.

References

[1] A. Publishing, *Radiation Oncology Physics a Handbook for Teachers and Students: Handbook of Radiation Oncology*. Independently published, 2020.

[2] E. S. Lebow, M. R. Bussière, and H. A. Shih, "Introduction to radiation therapy," in *Neuro-Oncology for the Clinical Neurologist*. Elsevier, 2021, pp. 28–37. [Online]. Available: https://doi.org/10.1016%2Fb978-0-323-69494-0.00003-8

[3] A. Torgovnick and B. Schumacher, "Dna repair mechanisms in cancer development and therapy," *Frontiers in genetics*, vol. 6, p. 157, 2015.

[4] J. W. H. Leer, "What the clinician wants to know: radiation oncology perspective," *Cancer Imaging*, vol. 5, no. special issue A, pp. S1–S2, 2005. [Online]. Available: https://doi.org/10.1102%2F1470-7330.2005.0027

[5] N. Tunçel, "Adaptive radiotherapy from past to future frontiers," *International Journal of Radiology &amp Radiation Therapy*, vol. 8, no. 2, pp. 81–84, may 2021. [Online]. Available: https://doi.org/10.15406%2Fijrrt.2021.08.00298

[6] D. Yan, F. Vicini, J. Wong, and A. Martinez, "Adaptive radiation therapy," *Physics in Medicine and Biology*, vol. 42, no. 1, pp. 123–132, jan 1997. [Online]. Available: https://doi.org/10.1088%2F0031-9155%2F42%2F1%2F008

[7] P. J. Keall, A. Hsu, and L. Xing, "Image-guided adaptive radiotherapy," in *Leibel and Phillips Textbook of Radiation Oncology*. Elsevier, 2010, pp. 213–223. [Online]. Available: https://doi.org/10.1016%2Fb978-1-4160-5897-7.00012-3

[8] C. E. Cardenas, J. Yang, B. M. Anderson, L. E. Court, and K. B. Brock, "Advances in auto-segmentation," *Seminars in Radiation Oncology*, vol. 29, no. 3, pp. 185–197, jul 2019. [Online]. Available: https://doi.org/10.1016%2Fj.semradonc.2019.02.001

[9] K. K. Brock, S. Mutic, T. R. McNutt, H. Li, and M. L. Kessler, "Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM radiation therapy committee task group no. 132," *Medical Physics*, vol. 44, no. 7, pp. e43–e76, may 2017. [Online]. Available: https://doi.org/10.1002%2Fmp.12256

[10] Z. Chen, W. King, R. Pearcey, M. Kerba, and W. J. Mackillop, "The relationship between waiting time for radiotherapy and clinical outcomes: A systematic review of the literature," *Radiotherapy and Oncology*, vol. 87, no. 1, pp. 3–16, apr 2008. [Online]. Available: https://doi.org/10.1016%2Fj.radonc.2007.11.016

[11] S. Oh and S. Kim, "Deformable image registration in radiation therapy," *Radiation Oncology Journal*, vol. 35, no. 2, pp. 101–111, jun 2017. [Online]. Available: https://doi.org/10.3857%2Froj.2017.00325

[12] N. Ziyan, "Deep Learning for Quality Assurance of Image Registration in Radiotherapy," *University of Manchester*, 2022.

[13] A. Maniyar, "CNN-based Quality Assurance of Image Registration in Radiotherapy," *University of Manchester*, 2022.

[14] L. Walker, "Quality Assurance of Image Registration in Radiotherapy using Deep Learning," *University of Manchester*, 2020.

[15] L. Fernández, "Deep Learning for Quality Assurance of Image Registration," *University of Manchester*, 2020.

[16] S. Bridger, "Deep Learning Deformable Image Registration Quality Assurance Summer 2021: A Summary of Findings," *University of Manchester*, 2021.

[17] Z. Chen, W. King, R. Pearcey, M. Kerba, and W. J. Mackillop, "The relationship between waiting time for radiotherapy and clinical outcomes: A systematic review of the literature," *Radiotherapy and Oncology*, vol. 87, no. 1, pp. 3–16, apr 2008. [Online]. Available: https://doi.org/10.1016%2Fj.radonc.2007.11.016

[18] D. Rueckert and J. A. Schnabel, "Medical image registration," in *Biomedical Image Processing*. Springer Berlin Heidelberg, 2010, pp. 131–154. [Online]. Available: https://doi.org/10.1007%2F978-3-642-15816-2_5

[19] A. A. Goshtasby, *2-D and 3-D Image Registration: for Medical, Remote Sensing, and Industrial Applications*. Wiley-Interscience, 2007.

[20] S. Arculeo, E. Miglietta, F. Nava, A. Morra, M. C. Leonardi, S. Comi, D. Ciardo, M. S. Fiore, M. A. Gerardi, M. Pepa, S. G. Gugliandolo, L. Livi, R. Orecchia, B. A. Jereczek-Fossa, and S. Dicuonzo, "The emerging role of radiation therapists in the contouring of organs at risk in radiotherapy: analysis of inter-observer variability with radiation oncologists for the chest and upper abdomen," *ecancermedicalscience*, vol. 14, jan 2020. [Online]. Available: https://doi.org/10.3332%2Fecancer.2020.996

[21] S. Broggi, E. Scalco, M. L. Belli, G. Logghe, D. Verellen, S. Moriconi, A. Chiara, A. Palmisano, R. Mellone, C. Fiorino, and G. Rizzo, "A comparative evaluation of 3 different free-form deformable image registration and contour propagation methods for head and neck MRI: The case of parotid changes during radiotherapy," *Technology in Cancer Research & Treatment*, vol. 16, no. 3, pp. 373–381, feb 2017. [Online]. Available: https://doi.org/10.1177%2F1533034617691408

[22] N. Kirby, C. Chuang, and J. Pouliot, "A two-dimensional deformable phantom for quantitatively verifying deformation algorithms," *Medical Physics*, vol. 38, no. 8, pp. 4583–4586, jul 2011. [Online]. Available: https://doi.org/10.1118%2F1.3597881

[23] T. Juang, S. Das, J. Adamovics, R. Benning, and M. Oldham, "On the need for comprehensive validation of deformable image registration, investigated with a novel 3-dimensional deformable dosimeter," *International Journal of Radiation Oncology Biology Physics*, vol. 87, no. 2, pp. 414–421, oct 2013.

[24] R. Varadhan, G. Karangelis, K. Krishnan, and S. Hui, "A framework for deformable image registration validation in radiotherapy clinical applications," *J. Appl. Clin. Med. Phys.*, vol. 14, no. 1, p. 4066, Jan. 2013.

[25] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, aug 2015. [Online]. Available: https://doi.org/10.1186%2Fs12880-015-0068-x

[26] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[27] A. Reinke, M. D. Tizabi, C. H. Sudre, M. Eisenmann, T. Rädsch, and M. Baumgartner, "Common limitations of image processing metrics: A picture story," 2021.

[28] H. Blumberg, "Hausdorff's Grundzüge der Mengenlehre," *Bulletin of the American Mathematical Society*, vol. 27, no. 3, pp. 116–129, 1920.

[29] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993. [Online]. Available: https://doi.org/10.1109%2F34.232073

[30] S. Ghosh, N. Das, I. Das, and U. Maulik, "Understanding deep learning techniques for image segmentation," *ACM Computing Surveys*, vol. 52, no. 4, pp. 1–35, jul 2020. [Online]. Available: https://doi.org/10.1145%2F3329784

[31] J. S. E. Hvitfeldt, *Supervised Machine Learning for Text Analysis in R*, 1st ed. New York, USA: Chapman and Hall, 2021.

[32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. London: MIT Press, 2016.

[33] V. M. S. Raschka, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*, 2nd ed. Birmingham, UK: Packt Publishing, 2017.

[34] E. Charniak, *An introduction to deep learning.* Boston, MA, USA: Addison-Wesley Educational, 2018.

[35] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, oct 1986. [Online]. Available: https://doi.org/10.1038%2F323533a0

[36] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, apr 1980. [Online]. Available: https://doi.org/10.1007%2Fbf00344251

[37] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 582–596, may 2019. [Online]. Available: https://doi.org/10.1007%2Fs10278-019-00227-x

[38] A. Khellal, H. Ma, and Q. Fei, "Convolutional neural network based on extreme learning machine for maritime ships recognition in infrared images," *Sensors*, vol. 18, no. 5, p. 1490, may 2018. [Online]. Available: https://doi.org/10.3390%2Fs18051490

[39] S. Skansi, *Introduction to deep learning: From logical calculus to artificial intelligence*, 1st ed. Cham, Switzerland: Cambridge University Press, 2018.

[40] J. Brownlee, *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch.* Melbourne, Australia: Jason Brownlee, 2016.

[41] Z. Gandomkar, P. L. Khong, A. Punch, and S. Lewis, "Using occlusion-based saliency maps to explain an artificial intelligence tool in lung cancer screening: Agreement between radiologists, labels, and visual prompts," *J. Digit. Imaging*, Apr. 2022.

[42] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[43] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, jun 2018. [Online]. Available: https://doi.org/10.1007%2Fs13244-018-0639-9

[44] W. R. Bosch, W. L. Straube, J. W. Matthews, and J. A. Purdy, "Data from Head-Neck_Cetuximab," 2015. [Online]. Available: http://doi.org/10.7937/K9/TCIA.2015.7AKGJUPZ

[45] A. LeNail, "NN-SVG: Publication-ready neural network architecture schematics," *Journal of Open Source Software*, vol. 4, no. 33, p. 747, jan 2019. [Online]. Available: https://doi.org/10.21105%2Fjoss.00747

[46] K. Pearson, "Vii. note on regression and inheritance in the case of two parents," *proceedings of the royal society of London*, vol. 58, no. 347-352, pp. 240–242, 1895.