PROSTATE CANCER DETECTION FROM MPMRI IMAGES

Parthiv Naidu

Supervised by:
Yipei Wang
Yipeng Hu

MPHY0012 BSc Final Report

Department of Medical Physics and Biomedical Engineering

Contents

Abstract

Background: Prostate cancer is one of the most common cancers in men worldwide (*Cancer in Men: Prostate Cancer Is #1 for 118 Countries Globally*, n.d.). However, the current screening tools are imperfect. For example, PSA testing has a false-positive rate of nearly 70% (Group & urgeinteractive, n.d.), leading to unnecessary biopsies (Sun et al., 2023a), increased infection risk and wasted time and money. Multiparametric MRI (mpMRI) has emerged as a superior method of imaging over MRI, improving detection of clinically significant prostate tumours. Yet, mpMRI interpretation is subjective and studies have shown substantial inter-reader variability in the identification and delineation of prostate lesions; this variability limits the full potential of mpMRI. This research explores whether an automated approach to delineating prostate lesions from mpMRI images is feasible. Specifically, we investigate if a U-Net convolutional neural network model could serve as an effective tool to enhance clinical prostate cancer imaging.

Methods and Results: A 3D U-Net (CNN) model was trained to annotated mpMRI scans (T2-weighted, ADC, and high b-value DWI sequences) of prostate cancer patients. Model performance was evaluated with the Dice Similarity Coefficient (DSC) primarily. Segmentation performance of the U-Net showed a bimodal trend. Large, well-defined tumours in the peripheral zone were often identified with moderate success (DSC around 0.5), whereas smaller, subtle, or transition-zone lesions were frequently missed (DSC ≈0.2 or even 0 for very faint foci). On average for all test samples, mean DSC was 0.3145, but increased to approximately 0.50 after excluding failure cases (cases when the lesion was not at all detected). This performance is comparable with the agreement range reported between multiple radiologists marking prostate lesions (Penzkofer, 2024).

Conclusions: The experiment demonstrates that a basic U-Net can in fact segment prostate cancer lesions on mpMRI at a level of performance equivalent to a human reader. Clinically, this kind of AI tool would be a valuable "second reader" or triage system – automatically flagging suspicious lesions for review by a radiologist (Sun et al., 2023). This could help reduce missed tumours, decrease inter-observer variation, and streamline radiology workflows. Importantly, the U-Net is not a replacement for expert judgment but rather a means to augment it by providing a consistent baseline suggestion. We acknowledge the model's limitations, particularly in detecting small or multifocal tumours, but these findings establish a promising baseline.

Future Work: Additional work will explore more advanced architectures (e.g. using attention mechanisms or transformer networks with the capability of global context capture (Kayalibay et al., 2017)), and alternative loss functions to improve training convergence and lesion boundary accuracy (Montazerolghaem et al., 2023). We also aim to include healthy patient scans within the training data to better depict normal anatomy and reduce false positives, ultimately increasing the model robustness for useful clinical application.

Introduction and Background

Context of the study

Prostate cancer is a major global health concern and one of the most prevalent malignancies among men. It is the most commonly diagnosed cancer in males, with around 1 in 8 men diagnosed with it over their lifetime. In the United Kingdom, prostate cancer accounts for 28% of all new cancer cases in men, resulting in 55,000 cases each year (*Key Statistics for Prostate Cancer | Prostate Cancer Facts*, n.d.). The incidence rates are similar across the globe, making it the 2nd most common cancer in men worldwide, with over 1.4 million new cases in 2022(Bergengren et al., 2023), and over 375,000 deaths worldwide in 2020(Lin et al., 2023). These figures motivate this research project. The sheer magnitude of the disease highlights the need for effective diagnostic strategies to successfully detect the cancer and eliminate uncertainty.

Diagnosing prostate cancer presents notable challenges. Early prostate cancer often has no symptoms and due to the silent and on-going nature of the disease, the cancer is typically detected after the onset of symptoms – such as difficulty urinating, frequent urination and blood in urine or semen. The pathway after a patient describes these symptoms begins with screening tests for prostate-specific antigen (PSA) levels and the digital rectal exam (DRE). PSA testing has improved early detection, however the method lacks the specificity for cancer, as men with elevated PSA do not always have cancer – they also indicate conditions such as prostate enlargement or inflammation. Therefore the high false positive rate of this test leads to unnecessary biopsies (David MK & Leslie SW, 2024)and anxiety. DRE testing involves a physician inserting a gloved finger into the rectum to feel the prostate; the physician is alerted to prostate cancer via a hard, lumpy or irregular prostate. DRE misses many cases of cancer, and whilst DRE palpates tumours sometimes, it can also miss them completely, especially when the tumour is small or located anteriorly (on the opposite side of the prostate to the physicians finger), making it an inconsistent standalone test(Lin et al., 2023). Consequently, men with abnormal PSA or DRE and commonly referred to a transrectal ultrasound-guided (TRUS) biopsy, which is the standard for obtaining a definitive tissue diagnosis. TRUS biopsy involves sampling the prostate with needles – they are invasive and carry risks such as bleeding, pain and infection. Therefore a TRUS biopsy informed only from PSA and DRE testing is not ideal, leading to many negative or unnecessary biopsy procedures(Ahmed et al., 2017), whilst also having the possibility of failure of detecting some significant cancers.

In recent years, multiparametric MRI (mpMRI) of the prostate has emerged as a better diagnostic tool which address the shortcomings of PSA and DRE testing. mpMRI combines anatomical imaging (T2-weighted MRI) with functional imaging techniques such as diffusion weighted MRI (DWI) to better visualise prostate tumours, which are all different imaging techniques that provides more information about the prostate(Yan et al., 2023). mpMRI is advantageous over normal MRI due to its ability to highlight suspicious lesions that may harbour clinically significant cancer, thus allowing for more

targeted biopsies. Additionally, if the scan is reassuring enough to radiologists and doctors, a biopsy can even be avoided. Evidence from the PROMIS study (Prostate MR Imaging Study), a large clinical trial, has proven the value of mpMRI. A standard TRUS biopsy showed a sensitivity of ~48%(Ahmed et al., 2017), whilst mpMRI showed a sensitivity of 93% for clinical significant cancers. In practical terms, an initial mpMRI can rule out high-grade disease in many patients, so that those with a negative scan might safely avoid an immediate biopsy. According to this trial, implementing mpMRI first could allow about 27% of men to avoid unnecessary biopsies, and reduce detection of harmless low-grade tumours by about 5%, while improving detection of significant cancer by 18% when MRI findings guide the biopsy(Ahmed et al., 2017). These findings have led to changes in clinical practice, with mpMRI being recommended before the first prostate biopsy for men with elevated PSA, in many guidelines. mpMRI is useful in directing targeted biopsies to suspicious areas, increasing the yield of significant cancer and reducing the number of random sampling errors. mpMRI also tends to preferentially detect higher grade tumours and can miss low-grade lesions, functionally acting as a filter for clinically important disease.

The mechanism radiologists use for MRI interpretation is called the Prostate Imaging – Reporting and Data System (PI-RADS), which provides scoring criteria for the tumours found from mpMRI scans. It defines how to evaluate and combine the different MRI sequences to estimate the likelihood of clinically significant cancer. Overall, mpMRI offers a more accurate, image guided approach, that can reduce overdiagnosis and improve the detection rates of aggressive cancers, addressing the limitations of PSA, DRE and TRUS biopsy.

mpMRI comes with downsides that this project aims to address also. Interpreting mpMRI is a complex task that require substantial experience as one radiologist assesses multiple imaging sequences and must use this information to create an overall assessment. The process comes with reader variability, and different radiologist will disagree on different PI-RADS scores for the same lesion. For example, when using PI-RADS v2, readers achieved a Cohen's kappa of around 0.55 for identifying clinically significant lesions (PI-RADS≥4), indicating moderate reproducibility (Rosenkrantz, Ginocchio, et al., 2016). Agreement is particularly lower for lesions in certain regions of the prostate – one study found interobserver κ values in the transition zone lesions were only ~0.39–0.51, compared to ~0.53–0.59 in the peripheral zone (Rosenkrantz, Ginocchio, et al., 2016) . Additional things that mislead the interpretation of mpMRI include post biopsy bleeding mimicking cancer, image artefacts or anatomical variation. These challenges mean that mpMRI, while powerful, has a subjective component and learning curve. A and the integration of deep learning techniques becomes highly pertinent in this scenario, as it fuels the motivation to explore methods of assisting radiologists and ensuring the full potential of mpMRI is reached in clinicals settings.

Multi-Parametric MRI and Imaging Modalities

Compared to single-sequence or "standard" MRI, which usually relies on T2W images alone, mpMRI combines structural and functional data, identifying subtle changes in tissue microstructure and vascular properties (Yan et al., 2023). The multiparametric technique markedly enhances sensitivity, identifying smaller or less obvious lesions, and specificity (e.g., separating benign from malignant features), and therefore maximizes the precise localization and delineation of prostate tumours (Ahmed et al., 2017). By simultaneous evaluation of morphologic, diffusion, and contrast-enhancement characteristics, mpMRI provides an integrated assessment that is not possible with routine MRI—ultimately reducing unnecessary biopsies and allowing for better clinical decision-making. ((Kasivisvanathan et al., 2018; National Institute for Health and Care Excellence, 2019).

Multiparametric MRI (mpMRI) of the prostate integrates three significant sequences—T2-weighted (T2W), diffusion-weighted imaging (DWI) with an apparent diffusion coefficient (ADC) map, and dynamic contrast-enhanced (DCE) imaging—to provide both anatomical and functional data (Weinreb et al., 2016). T2W imaging gives a high-resolution structural image, where normal peripheral zone tissue is bright and tumours typically are darker (hypointense) regions (Turkbey et al., 2019). DWI/ADC quantifies the motion of water protons, highlighting malignant masses that restrict diffusion—these appear hyperintense on DWI with reduced ADC values (Chen et al., 2021). DCE imaging, on the other hand, involves the utilization of a contrast agent and monitoring vascular patterns of enhancement; areas of enhanced vascularity and early enhancement are more likely to represent clinically relevant malignancies (H. Dickinson et al., 2013a). Each modality addresses a distinct feature of tumour pathology—T2W (anatomical architecture), DWI/ADC (cellularity), and DCE (vascularity)—and thus collectively improves lesion detection and characterization (Weinreb et al., 2016).

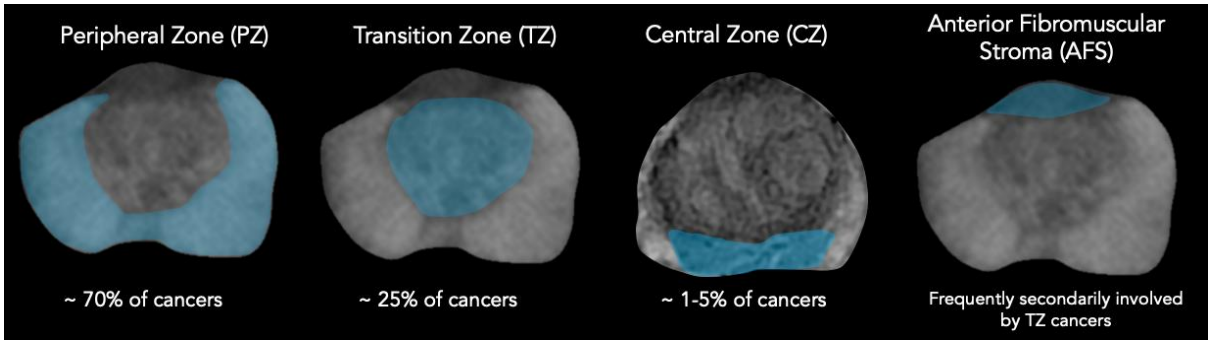

*Figure 1 - Diagram showing the peripheral, transition and central zones of the prostate (*Prostate MRI - RAD-ASSIST*, n.d.)*

Importance of AI and Deep Learning in Prostate Cancer Imaging

Artificial intelligence and deep learning techniques have developed rapidly in medical image analysis and offer promising solutions to reduce the variability and inefficiencies of prostate cancer diagnosis. An AI system can be used as a second pair of eyes that can interpret images consistently in the same way, depending on a learned pattern of

prostate cancer. The power of this technology is its ability to reduce human error and decrease inter-observer variation. A learned framework will add standardisation, providing the same prediction for identical output. As there are such a high number of prostate MRIs being undertaken, AI is able to augment diagnostic efficiency through flagging regions of suspicion to radiologist reporting by streamlining the interpretation. This cuts down on the time that radiologists spend per case and help focus attention on cases that have a higher clinical significance of cancer. Research indicates that the automation of imaging analysis can improve both accuracy and efficiency of the diagnostic process – it can enhance the consistency of tumour contours and the speed of planning radiotherapy treatments (Lucido et al., 2023). Additionally, AI that marks cancerous regions ensures that small but important lesions aren't overlooked, streamlining the radiology workflow. Early research indicates that such automation can indeed improve diagnostic workflow: one study noted that an AI-based second-reader system for prostate MRI outperformed over 70% of general radiologists in detecting prostate cancer on MRI (Alqahtani, 2024; Cao et al., 2021). Another investigation showed an AI could achieve a high sensitivity (around 85–87%) for detecting clinically significant cancers (e.g., PI-RADS 4–5 lesions or Gleason ≥7 tumours), comparable to experienced human readers (Alqahtani, 2024). Indeed, the automation of imaging analysis has been shown to improve both accuracy and efficiency – for instance, deep learning models have demonstrated segmentation performance on prostate MRI that is on par with expert radiologists for delineating the prostate gland (Fassia et al., 2024).

The current applications of deep learning in prostate MRI involve segmenting lesions or detecting lesions. For segmentation based approaches, models work by producing a voxel level mask that outline the regions of the image likely to contain cancer. In detection-based models, the model typically identifies suspect regions through bounding boxes or classifying outputs (Litjens et al., 2017) representing zones with the probability of hosting malignancies. Both methods are valuable towards identifying prostate cancer but segment-based techniques prevail in providing closer delineation of lesion edges vital in guiding biopsies as well as in planning treatments (L. Chen et al., 2021). Of all available architectures, transformer-based networks (e.g., TransUNet) have the capability of accepting long-distance dependencies but typically use more large-sized datasets as well as larger computation resources (L. Chen et al., 2021). As opposed to this, the more basic convolutional structure U-Net has demonstrated reliably good performance even on quite small medical image data sets because of its encoder-decoder structure and skip connections. Therefore, this project employed the U-Net rather than a vision transformer because it has worked well in practice for biomedical segmentation, is eminently suited for few-data scenarios, and has relatively less complexity. Therefore, a more pragmatic, reliable choice to employ for segmenting prostate MRI is U-Net (Isensee et al., 2021).

Convolutional neural networks (CNN) have been the backbone of deep learning models. The most common CNN architecture for biomedical image segmentation tasks is known as U-Net. Introduced in (Ronneberger et al., 2021), U-NET feature an encoder-decoder structure with skip connections that enable precise localisation by combining

coarse semantic information with fine spatial details. The architecture has achieved start of the art performance in medical image segmentation in other fields. For instance, U-Net variants have been used to segment brain tumours on MRI with high accuracy (Yousef et al., 2023). Reserachers have explored various improvements such as optimised loss functions to further enhance performance. The appeal of U-Net for this segmentation task is its ability to learn from relatively modest sized datasets and produce pixel wise predictions that align well with radiologist annotations.

The second important aspect of using AI for mpMRI is the combination of the multi-parametric data. Radiologists would traditionally view T2-weighted, DWI (traditionally depicted by the ADC map), and mentally combine this information based on the PI-RADS scoring rules to form a global opinion (Turkbey et al., 2019). Replicating this process is not trivial for AI. Some models utilize the multiparametric images as multi-channel input to one network (essentially permitting the network to learn how to integrate them), while others have attempted to precisely imitate the radiologist's decision-making rules. For example, Combiner networks (Yan et al., 2023) mimic the process of scoring every modality of an MRI and then applying a rule-based combination so as to make predictions regarding the probability of cancer. Their approach introduced a "HyperCombiner" network which can be conditioned on different modality-weighting rules at inference time, and it can implicitly learn the optimal way of fusing information from T2W, DWI, and DCE while maintaining accuracy. It points up the key principle that good prostate cancer AI has to leverage all available MRI information harmoniously since different tumours might best be imaged on different sequences, like some on DWI, some on DCE, etc (H. Dickinson et al., 2013b). Deep learning systems, through having learned on extensive datasets with well-documented outcome values, potentially will be able to detect weak trends over sequences difficult for humans to combine into one picture time and again consistently.

Notably, the motivation for AI in prostate cancer imaging stems from a clinical need for automated, standardized interpretation. A properly trained AI system could help ensure that all mpMRIs are reviewed to the same degree and with the same level of quality, eliminating the issue of inter-reader variability in PI-RADS scoring (Rosenkrantz, Verma, et al., 2016). It may also help with triage testing in busy practices – e.g., quickly differentiating plain normal testing (no suspicious lesion) from plain high-grade cancer, so radiologists can focus on high yield cases (Twilt et al., 2021). Additionally, AI could provide decision support borderline cases (e.g., PI-RADS 3 lesions) by integrating weak imaging features that are predictive of cancer and ideally increasing specificity such that fewer men with benign conditions are sent to biopsy (Kasivisvanathan et al., 2018). Early studies have shown that deep learning models can perform as well as expert radiologists for prostate cancer detection on MRI (Seetharaman et al., 2021), but validation and calibration are needed (Oktay et al., 2018a).

It's also crucial that the way we evaluate AI performance is considered – a model that gets a high Dice similarity for segmentation can still not detect small isolated tumour foci that are of clinical importance. Scientists have pointed out that traditional voxel-

level metrics (e.g., Dice or Hausdorff distance) are not necessarily consistent with clinical detection outcomes. Consider an algorithm is segmenting one large tumour perfectly but overlooks a second target. Then the Dice measure can still be high despite having missed one cancer. Therefore, more recent research advises the efficacy metrics at the lesion level (precision and sensitivity per lesion) as well as voxel metrics to assess an algorithm's actual clinical utility. AI can assist in revolutionizing prostate cancer detection on mpMRI by improving consistency, accuracy, and efficiency – but it is essential to test these tools rigorously in clinically meaningful ways.

Challenges in Prostate Cancer Detection

This section will outline the several key challenges in automating prostate cancer detection from mpMRI images with deep learning techniques:

Prostate cancer tumours do not have a single obvious appearance, and they manifest differently across T2W, DWI and ADC sequences. Given the tumours intricate and detailed multi-model features, the model must learn a combination of texture and intensity patterns, across these three sequences in order to reliably detect the lesion. For a CNN architecture with limited depth, the network may not be sufficiently expressive and the detail of the tumours may be hard to capture. Ensuring the model effectively integrates information from T2, DWI, and possibly DCE is a non-trivial challenge (and as discussed, is an active area of research with approaches like combiner networks addressing this integration (Yan et al., 2023).

Additionally, prostate cancer lesions vary in size, shape and number, from large obvious masses to multiple medium masses to small barely perceptible lesions (H. Dickinson et al., 2013b). The U-Net model may be biased towards detecting the most common kind of training data (larger and higher contrast lesions), at the expense of a decreased sensitivity to small tumours, where smaller lesions with fainter contrast can be missed. The model may also fail to detect leisons from patients with multiple lesions, which is a common occurrence in models of simpler architecture – this can lower segmentation metrics like the dice score coefficient. Therefore the challenge lies in handling the full spectrum of possible lesions.

Even when a lesion is detected, precisely delineating its boundaries is challenging. The transition between tumour and healthy tissue on MRI can be gradual or obscured by imaging noise. This is especially true in cases of lower contrast or lesions in uncommon locations (e.g., where tumour tissue may blend with normal tissue). The apex (the bottom tip of the gland) is notably difficult for segmentation because the gland thins out and image slices often include partial volume effects, making it hard to say exactly where the tumour ends (Milletari et al., 2016). As a result, segmentation algorithms often produce rough or “blobby” lesion masks that do not capture the fine irregular margins that a radiologist might draw. Achieving high boundary precision is difficult; it may require higher-resolution modelling or special loss functions that emphasize contour accuracy. Ambiguous prostate lesion boundaries especially towards the apex region have been cited as a reason why prostate MRI segmentation is inherently tough.

Even human experts have variation here (radiologist segmentations of the same tumour can differ considerably at the edges, contributing to interobserver DSC values in the 0.5–0.8 range for tumour masks (M. Y. Chen et al., 2020).

Performance can differ by prostate region. Lesions in the apex and base regions are harder to detect due to prostate gland's shape and MRI slice orientation (Milletari et al., 2016). Tumours in the transition zone (central gland) can be obscured by benign prostatic hyperplasia nodules -small swellings in the nodules, leading to an enlarged prostate - making them difficult for both radiologists and AI to distinguish from benign tissue. In fact, PI-RADS scoring rules designate T2-weighted imaging as the primary determinant for transition zone lesion suspicion, since cancer there typically appears as an irregular, infiltrative low-signal area on T2, but this can be subtle (Turkbey et al., 2019). Whilst some research has been done to explore the training separate models for different zones (Bhayana et al., 2021), these region-specific challenges mean an AI might perform well in one area of the prostate but consistently underperform in others, reducing overall reliability.

Research Questions and Objectives

Given the above context, this research project is motivated by the need to utilise deep learning techniques for improving prostate cancer detection from mpMRI images. The aim of the research is to evaluate the performance of a CNN-based U-Net model for the segmentation of prostate cancer lesions. The U-Net model was chosen due to its proven success in biomedical image segmentation tasks, especially on limited datasets, and its architecture which preserves fine spatial details critical for accurately delineating prostate cancer lesions (addressed in methodology). By applying this model to prostate mpMRI, this study intends to assess how well a relatively simple deep learning model can delineate and segment cancerous lesion against the complex background of normal anatomy and benign findings.

To achieve this, the research project sets out the following specific objectives:

1. **How accurately does a simple CNN U-Net segment prostate cancer lesions in multiparametric MRI scans?**

This question examines the core performance of the model, by answering whether the model is able to correctly delineate cancerous lesions and how its accuracy compares to reference standards (radiologist annotations). This entails training and testing the CNN U-Net model on mpMRI data to quantify its accuracy in segmenting the lesions.

2. **What are the key limitations of using a simple U-Net for prostate cancer segmentation?**

This question involves analysing the cases where the model underperforms or fails. These cases include the model missing a lesion or falsely segmenting normal tissue, in order to fully understand the shortcomings of using a basic U-Net architecture. This

objective also includes identifying patterns to the errors – for example if a falsely identified lesion is due to a motion artifact in the MRI image, and considering in what circumstances does the U-Net face difficulties. This research question will highlight areas of improvement and be the basis of identifying needs of more advanced architectures.

3. **Can the diagnostic workflow in clinical settings be made more efficiently by incorporating AI-based segmentation, without compromising accuracy?**

This helps in investigating how the integration of AI segmentation tools could impact clinical workflows and diagnostic efficiency. This involves evaluating whether/to what extent does using the U-Net's output could speed up image interpretation, or if the model has potential to reduce the number of unnecessary biopsies by improving confidence in imaging results. A key component of this question is whether any gains in efficiency or consistency can be achieved without compromising diagnostic accuracy. Will the model be able to maintain high sensitivity and specificity? Will it be able to perform under these metrics, while being practical tool with the ability to alleviate the workload and decrease the variability associated in diagnosis?

By thoroughly examining these questions, this research project aims to provide evidence on the effectiveness of deep learning-based automation in prostate MRI analysis. It also seeks to offer insights into whether a relatively simple model like U-Net is adequate or where it may fall short.

Methods and Experiments

Dataset and Preprocessing

This study utilised a large mpMRI dataset of prostate cancer cases, sourced from several clinical trials at University College London, to train and evaluate the deep learning model. A total of 851 patient's mpMRI scans are included in the overall dataset; the individual clinical trials that make up the dataset are: SmartTarget(Yan et al., 2023), PICTURE (Simmons et al., 2018), ProRAFT (Orczyk et al., 2021), Index (L. Dickinson et al., 2013)and PROMIS ((El-Shater Bosaily et al., 2015), PROMIS Study Dataset - Open Access Request - NCITA, n.d.). Each patient's MRI consists of 3 imaging sequences: a T2-weighted image, an apparent diffusion coefficient (ADC) map and a high b-value diffusion-weighted image (DWI). All cases were annotated by expert radiologists, following PI-RADS guidelines, delineating each lesion and these annotations serve as ground truth segmentation masks for training.

The dataset is made up of raw MRI scans. Consequently, data preprocessing steps were taken to ensure data consistency and to prepare the data as inputs for the U-Net model.

The first pre-processing step was co-registration. This spatially aligns all these different datasets for a single patient. Even if acquired in the same session, slight movements or inherent geometric differences between scan types mean they aren't perfectly lined up

initially. Co-registration corrects for this by geometrically transforming the ADC, DWI, and lesion mask to match the spatial coordinate system of a chosen reference image (Modat et al., 2010).

Following co-registration, the datasets were resampled to a common spatial reference frame. This ensures that all images and the mask for a patient not only align spatially but also have the exact same digital grid structure (i.e., the same voxel size and matrix dimensions) (Sotiras et al., 2013). For this process, the script designated the T2-weighted image as the reference, meaning its spatial grid became the target grid for all other datasets.

Resampling entails calculating the appropriate value (image intensity for images, mask label for masks) for every voxel in this new common grid based on the values in the original (but co-registered) data. This is carried out by interpolation methods. The technique used has a major influence on maintaining the significance of the data:

- Linear Interpolation of ADC and DWI: Linear interpolation was done for the ADC and DWI images, which contain continuous or semi-continuous intensity values related to physical properties of the tissue. In this method, the intensity value of a new voxel is calculated as a distance-weighted average of the intensity values of its direct neighbours (typically 4 in 2D, 8 in 3D) in the original image. It produces relatively smooth results and is suitable for data where values change gradually, thus preserving the quantitative nature of the ADC and DWI signals during the grid transformation (Villanueva-Meyer et al., 2017).
- Nearest-Neighbour Interpolation for Lesion Masks: For the lesion masks, categorical data (e.g., a voxel is either '0' for background or '1' for lesion), nearest-neighbour interpolation was used. This method simply transfers the value of the nearest individual voxel from the original mask to the same voxel location in the new grid. This is important for masks since it avoids generating artificial in-between values (e.g., 0.5), such that all voxels in the resampled mask again have a valid, discrete label ('0' or '1'). This maintains the integrity of the binary segmentation boundary (Chareshneu et al., 2024).

Following co-registration and resampling, the images were processed to ensure precise registration between modalities and create quality control visualizations for verification. The voxel dimensions were standardised (on the order of 0.5x0.5x1 mm, typical for prostate MRI) so that T2, ADC, and DWI were spatially aligned. The 3D volumes were then sliced into 2D axial images for input. Each slice was padded to a fixed size (192x192 pixels) to provide a consistent input, focusing on the prostate region while removing excessive background. This ensured that the network's field of view included the entire prostate and immediate surroundings but not irrelevant regions of the pelvis.

The dataset was then divided into training, validation and test sets, with a respective ratio of 0.7:0.1:0.2 (Chareshneu et al., 2024). The dataset was shuffled into a random order, to ensure a similar distribution of lesion sizes across the 3 sets (Chareshneu et al., 2024). The test set remained untouched for the entire training process and was only

used for the final performance test. The validation set was used for tuning the model, monitoring dice score and to check the model's training process. The final input to the network was a set of 2D slices with 3 channels (T2W, ADC and DWI) and the target output of the model was a 2D binary mask of the same dimensions, indicating the predicted tumour.

Model Architecture

We implemented a U-Net-based convolutional neural network for prostate lesion segmentation. The U-Net is an encoder–decoder network with skip connections (see Figure 2) that was originally proposed for biomedical segmentation tasks. Our implementation follows the typical 2D U-Net architecture (Ronneberger et al., 2015). The model input is 256×256 in size with 3 channels (T2, ADC, DWI), and the output is a segmentation map of the same size but with one channel representing the cancer probability of each pixel.

The encoder pathway of the network consists of a series of convolutional layers and downsampling operations that progressively reduce the spatial dimensions while learning higher-level feature representations (Ronneberger et al., 2021). Each encoding step uses two 3×3 convolutions (with ReLU activation) followed by 2×2 max pooling to halve spatial dimensions. We used four pooling steps in the encoder (from 256×256 to 16×16), doubling the number of feature channels at each step (e.g., 32 after the first block, 64 after second, and so forth, to 512 channels at the bottom). This yields a bottleneck layer with drastically reduced spatial dimensions but rich feature density (Milletari et al., 2016).

The decoder pathway is symmetric to the encoder: at each level, an upsampling (2×2 transposed convolution) doubles the resolution, and the feature maps of the corresponding encoder are concatenated through skip connections. These skips carry high-resolution information from the contracting path that allow the expanding pathway to localize features that would otherwise be lost. After each concatenation, two 3×3 convolutions refine features. Feature channel counts are halved at each decoder step to reflect the symmetric architecture of the encoder. At the final layer, a 1×1 convolution produces a single-channel image, and a sigmoid activation is applied to obtain pixel-wise cancer probabilities (values between 0 and 1 at each pixel) (Litjens et al., 2017).

Our U-Net model was designed to be simple (no complicated modules like attention gates or residual blocks, and weights initialized from scratch) as a baseline. The simplicity reduced computational cost and overfitting risk, as well as our dataset size. While with a "plain" U-Net, the architecture incorporates the essential design required for segmentation: the combination of global context and local detail through the encoder-decoder and skip structure. We hypothesized that if even this basic U-Net worked reasonably well, it would validate the direction of our approach and indicate that more complex architectures would be capable of improving performance if needed. The relatively low parameter count also promises quicker training of the model and possibly better generalization on smaller datasets (Zhang et al., 2024).

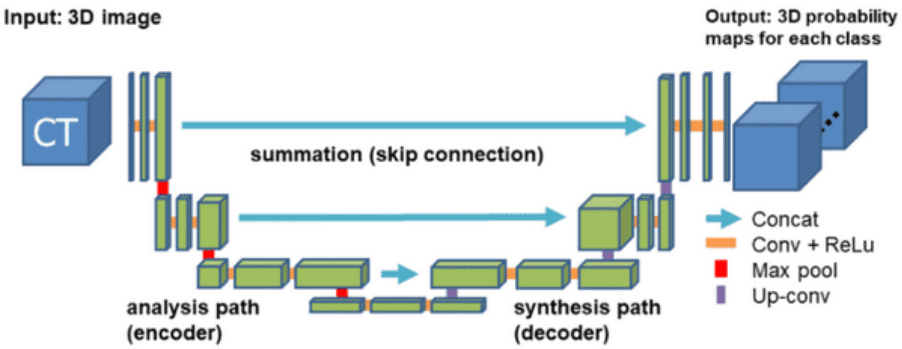

Figure 2 – Diagram showing U-Net structure, encoder, decoder and skip connections (Oktay et al., 2018b)

Loss Function Evaluation

Choosing a proper loss function is essential for medical image segmentation problems, especially when the problem involves extreme class imbalance, such as in the case of prostate lesion segmentation, where tumour areas take up only a minority of the overall image volume (Sarkar & Li, 2022). In this project, we used the Dice loss as the only objective function to train the U-Net model.

The Dice loss is a direct formulation of the Dice Similarity Coefficient (DSC), and it is used as the primary evaluation measure in this work.

$$\mathbf{Dice} = \frac{2|X \cap Y|}{|X| + |Y|}$$

where X is the estimated binary segmentation mask, and Y is the true ground mask. The Dice coefficient measures the overlap between the predicted and the true lesion area spatially, and its values are between 0 (no overlap) and 1 (optimal overlap).

In comparison to pixel-wise loss functions such as binary cross-entropy (BCE) that operate per voxel, Dice loss is considering the entire segmentation. That is why Dice loss is invariant to class imbalance between lesion and background voxels: predictions that omit small tumour regions (the common failure mode when training with BCE alone) are highly penalized, since omitting all of the lesion implies zero overlap and hence maximum Dice loss (Bakas et al., 2018a).

Although some studies recommend employing a combined loss function, e.g., BCE + Dice loss, to trade off local and global segmentation accuracy (e.g. Montazerolghaem et al., 2023), initial experiments showed that Dice loss alone was sufficient for training

stability and performance. In fact, we observed that employing Dice loss as a single objective yielded stable convergence, reasonable training speed, and proper sensitivity to small lesions. This aligns with previous research that shows Dice loss will be extremely good in biomedical segmentation tasks where lesion sizes are extremely heterogeneous and class imbalance is a gigantic problem (Ronneberger et al., 2015).

Briefly, Dice loss was chosen due to its suitability with the evaluation metric and its robustness to the class imbalance of prostate cancer segmentation. It encourages the model to maximize spatial overlap rather than voxel-wise accuracy, which is clinically more meaningful as the detection and delineation of lesion regions is the end objective (Kokkinos et al., 2019).

Model Training and Evaluation

The U-Net model was implemented in Python using PyTorch, a widely adopted deep learning framework (Paszke et al., 2019). The Adam optimiser was used with an initial learning rate of 0.001 (Kingma & Ba, 2015). When training a neural network, a gradient represents the change of the loss function with respect to the network's parameters, and has a direction and magnitude. The goal is to minimise the loss, so the model should move in the opposite direction of the gradient. The Adam optimiser achieves this by adaptively adjusting the learning rates for each parameter (weights and biases) based on the estimates of the prior two gradients. The use of the Adam optimiser meant that this learning rate provided a stable convergence (the point in training where the model's performance during training improves consistently and plateaus, showing the model has learned the underlying patterns; further training will be unlikely to yield significant improvements). Training the model will run for 100 epochs. To save energy from processing the results, the nevesaary code was included to stop the model after 10 epochs if there was no improvement in the loss score.

The weights are the learnable parameters of each convolutional layer (*GitHub - Janishar/Mit-Deep-Learning-Book-Pdf: MIT Deep Learning Book in PDF Format (Complete and Parts) by Ian Goodfellow, Yoshua Bengio and Aaron Courville*, n.d.). These parameters are organized as multi-dimensional tensors (arrays) that represent the convolutional kernels or filters. During the forward pass, these kernels convolve over the input image (or feature map), capturing local patterns such as edges, textures, or more complex features. Each convolutional block in both the contracting (encoder) and expanding (decoder) paths has its own set of weight parameters. Typically, weights are initialized randomly in a manner that ensures that each feature map starts with approximately the same variance. This careful initialization is essential for maintaining a stable flow of gradients during training (Glorot & Bengio, n.d.). Apart from the weights of the convolution, there is a bias term in each layer of the U-Net model. The bias is included in the result of the convolution operation such that each filter can be shifted. The shift is significant as it allows the network to learn features even when the input values are zero.

The main performance metric for guiding optimisation during training was the dice coefficient. The model was saved after every epoch of training, and parameter metrics were recorded for each epoch. After training, the epoch with the best parameters was chosen for validation. This model was then applied to the independent test set. For each test patient, the model processed the T2, ADC, DWI slices and output a probability map for each slice. The probability maps were thresholded at 0.5 to generate binary segmentation masks (M. Y. Chen et al., 2020).

The dice coefficient computed per patient by comparing the predicted mask with the ground truth mask across the entire 3D volume. If there were more than one lesion for a patient, all of them were employed in the DSC computation (so missing one lesion returns a low DSC for that patient). In order to calculate the dice score, we collected pixel-wise sensitivity (fraction of true tumour pixels detected) and specificity (fraction of non-tumour pixels identified correctly) for reference. By analysing these numerical metrics and taking a look at the segmentation outcome, we were able to estimate the model accuracy and identify failure modes. The entire experiment was conducted with valid separation of training/validation/test in order to supply an unbiased view (M. Y. Chen et al., 2020). The succeeding chapters discuss experimental results and results of this analysis.

Results

Quantitative and Qualitative Results

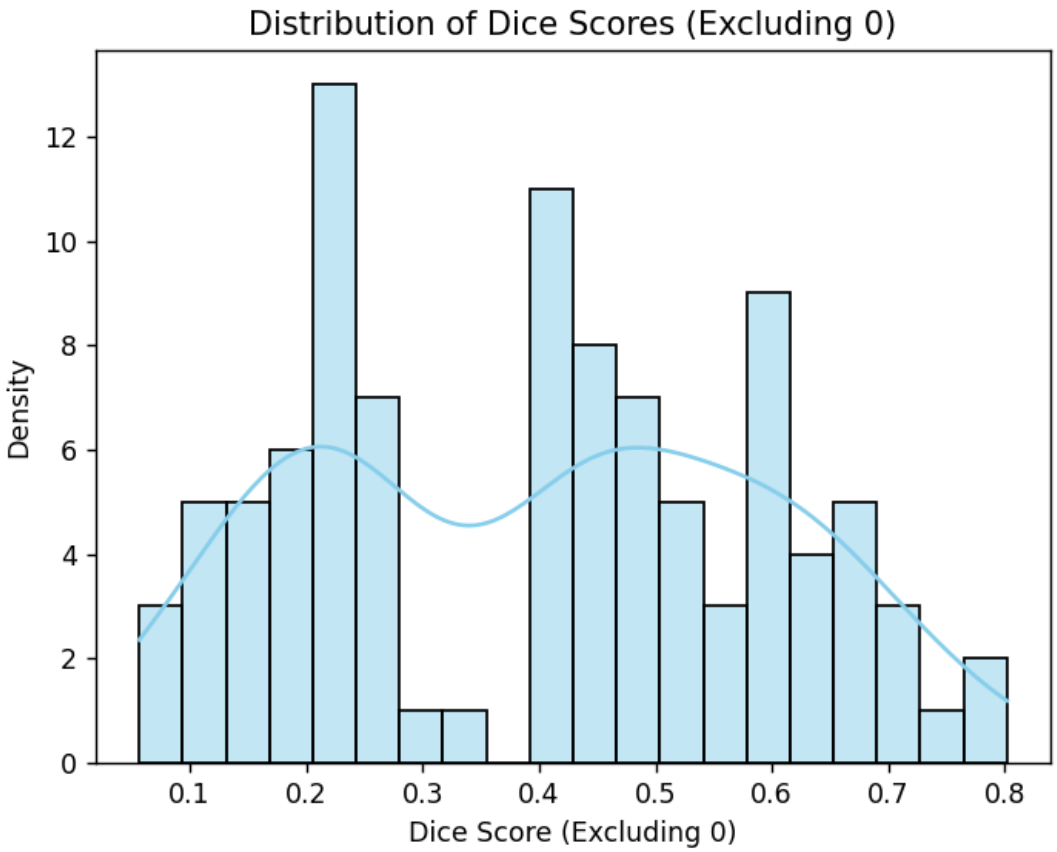

Figure 3 - Distribution of Dice scores from the test set

The dice score of the convolution neural network U-Net model indicates a moderate performance overall. The mean DSC achieved by the U-Net model was 0.3145 ±0.2370, with the median DSC = 0.4160.

A significant number of cases – approximately 20% of the test set - involved the model completely failing to identify the lesions, resulting in a DSC of 0. The mean DSC of the U-Net model without any 0 values was 0.5028 ±0.2085 and the median DSC was 0.62. Even with these zeros, the remainder of the level of agreement for good segmentations is comparable to the agreement between different human raters – earlier work reported inter-observer DSC of approximately 0.5 for prostate tumour segmentation on MRI. Therefore, although the overall mean is affected by the greater number of zero results, the model's accuracy on instances where it did find some overlap is within a reasonable range.

The distribution of dice scores from the test set of the U-Net model on 128 patients revealed a bimodal pattern(2 distinct peaks), characterised by 2 distinct peaks. The distribution featured two prominent modes: one lower mode around approximately a Dice score of 0.2, and one higher mode with Dice scores between approximately 0.5.

This indicates that the model's segmentation performance varied greatly throughout the test set, with two sets of cases having different performances.

Qualitative analysis of model outputs allowed us to categorize failure modes and better interpret the two modes in the Dice distribution:

- False Negatives (Missed Lesions) – see Figure 4,5,6 for examples

Most of the DSC = 0 cases were due to false negatives, i.e., the model output was zero lesion mask but one or more lesions were present in the ground truth. These failures were particularly prevalent in lesions at the base or apex of the prostate, which are anatomically challenging and tend to be plagued by reduced image contrast. In other instances, the lesions were very small or occurred subtly, and hence were hard for the model to identify with high confidence.

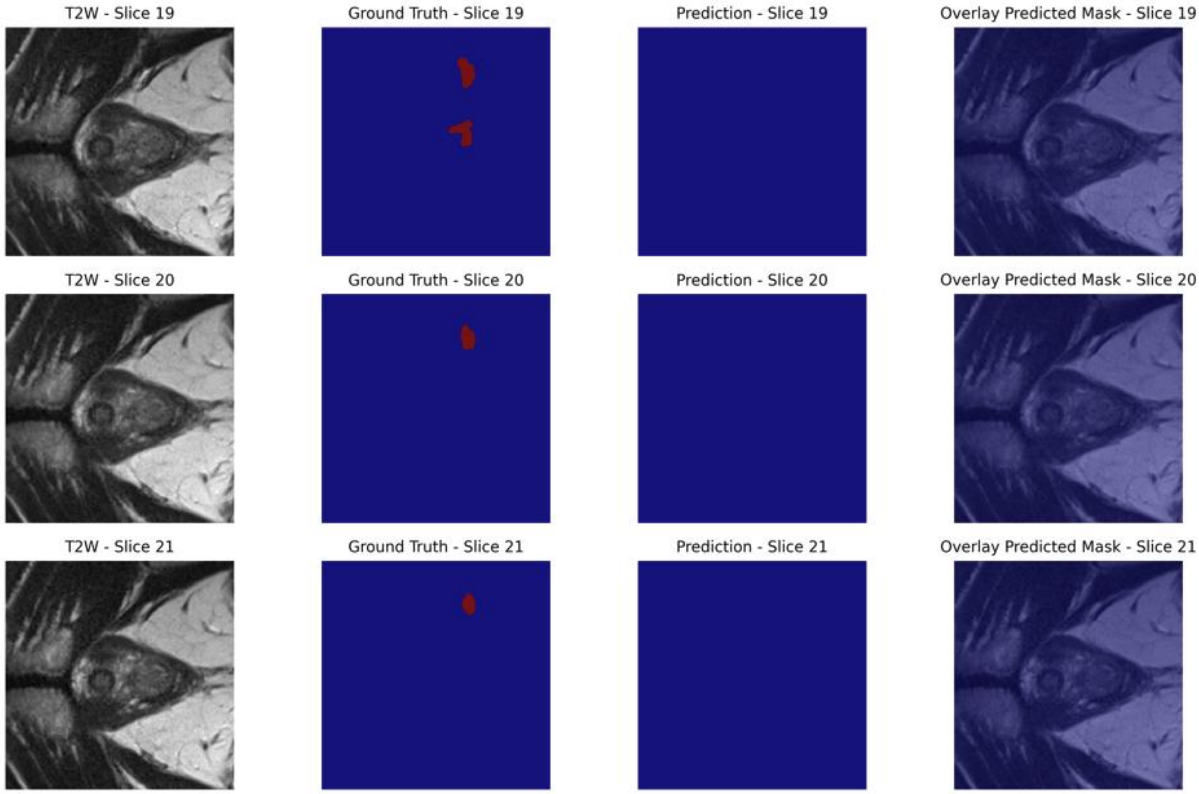

Figure 4 -visualising a test case where DSC = 0.0000. The model struggled to detect lesions towards the base/apex regions of the prostate.

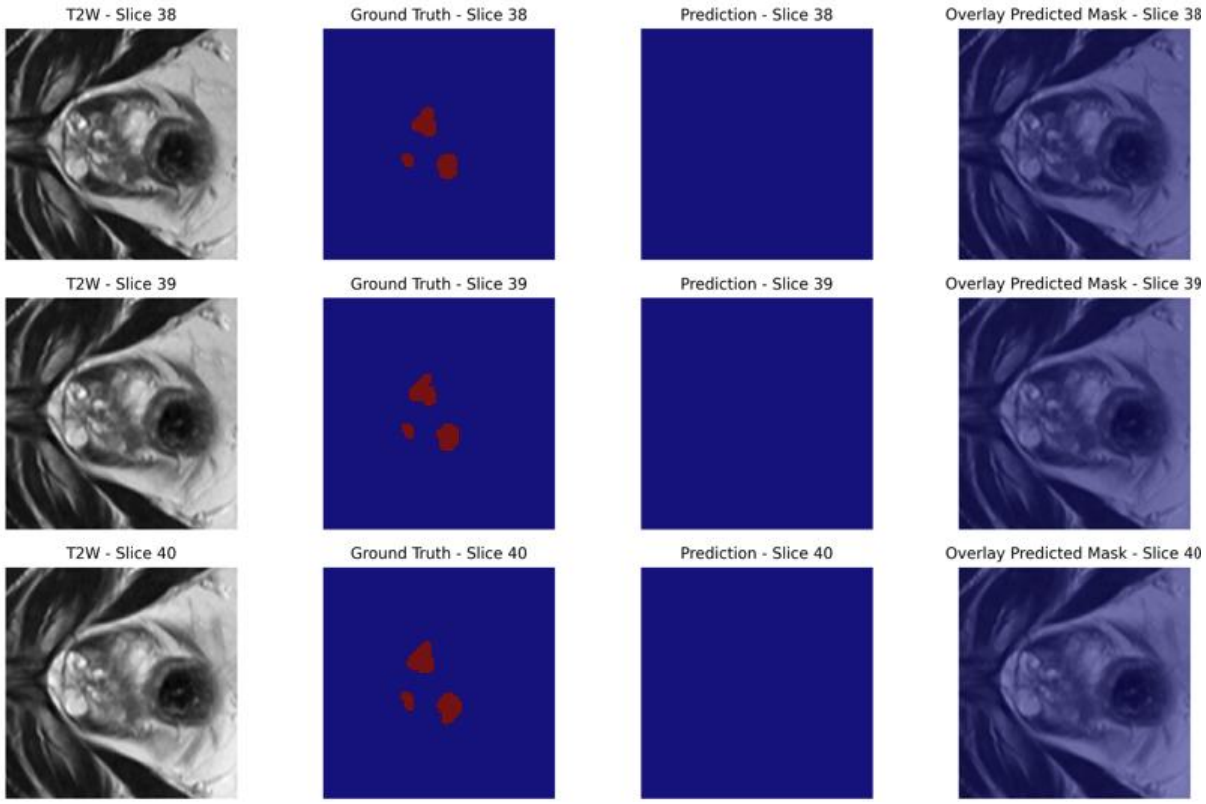

Figure 5 - Visualising a test case where DSC = 0.0000. The model struggled to detect lesions of multiple foci.

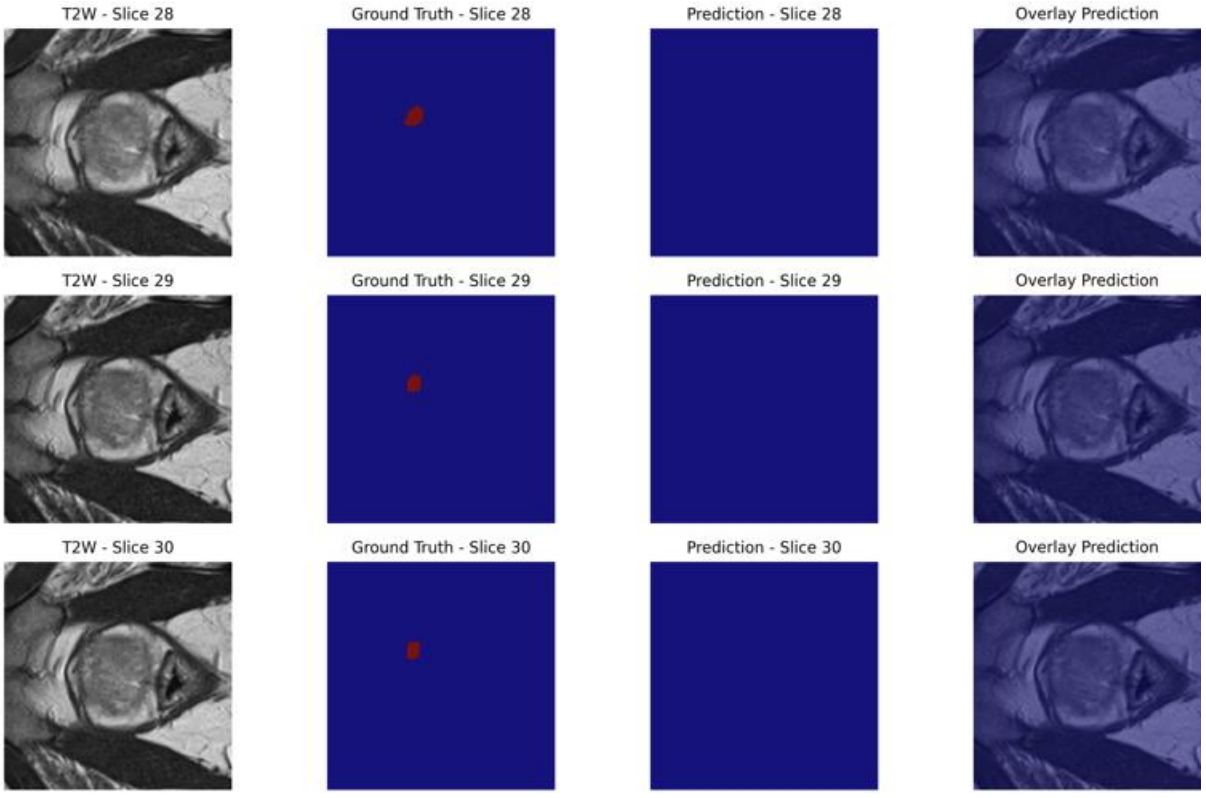

Figure 6 - Visualising a test case where DSC = 0.0000. The model struggled to detect particularly small/early stage lesions

- Partial Segmentations (Low Dice ≈ 0.2) – see Figure 7 for an example

At the lower peak of the distribution (~0.2 DSC), the model was likely to detect a lesion but not segment it well, capturing only a fraction of the lesion or one of several foci. This was particularly observed in multifocal disease, in which the model either captured only one lesion or detected isolated areas that did not significantly overlap with the ground truth. Additionally, the model would over predict lesions, detecting lesions that don't exist.

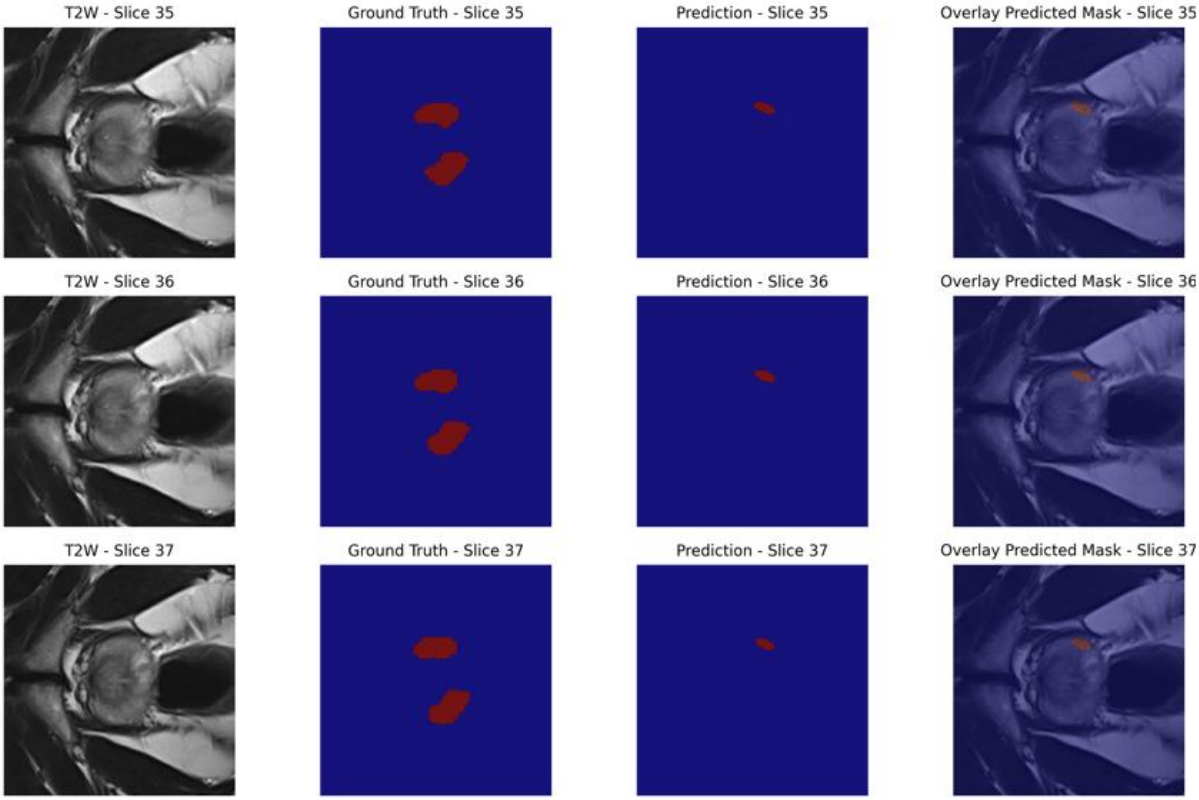

Figure 5 – Visualising a test case where DSC = 0.1719. The model struggled with identifying multiple lesions, predicting one foci out of the two

- False Positives (Incorrect Regions) – see Figure 8 for an example

Some of the low-scoring predictions (DSC ≈ 0.2) were due to false positives, where the model predicted lesions outside the ground truth region, mostly along the edge of the prostate. These tended to be near benign structures or motion artefacts, which can have appearance patterns close to true lesions on DWI or ADC maps.

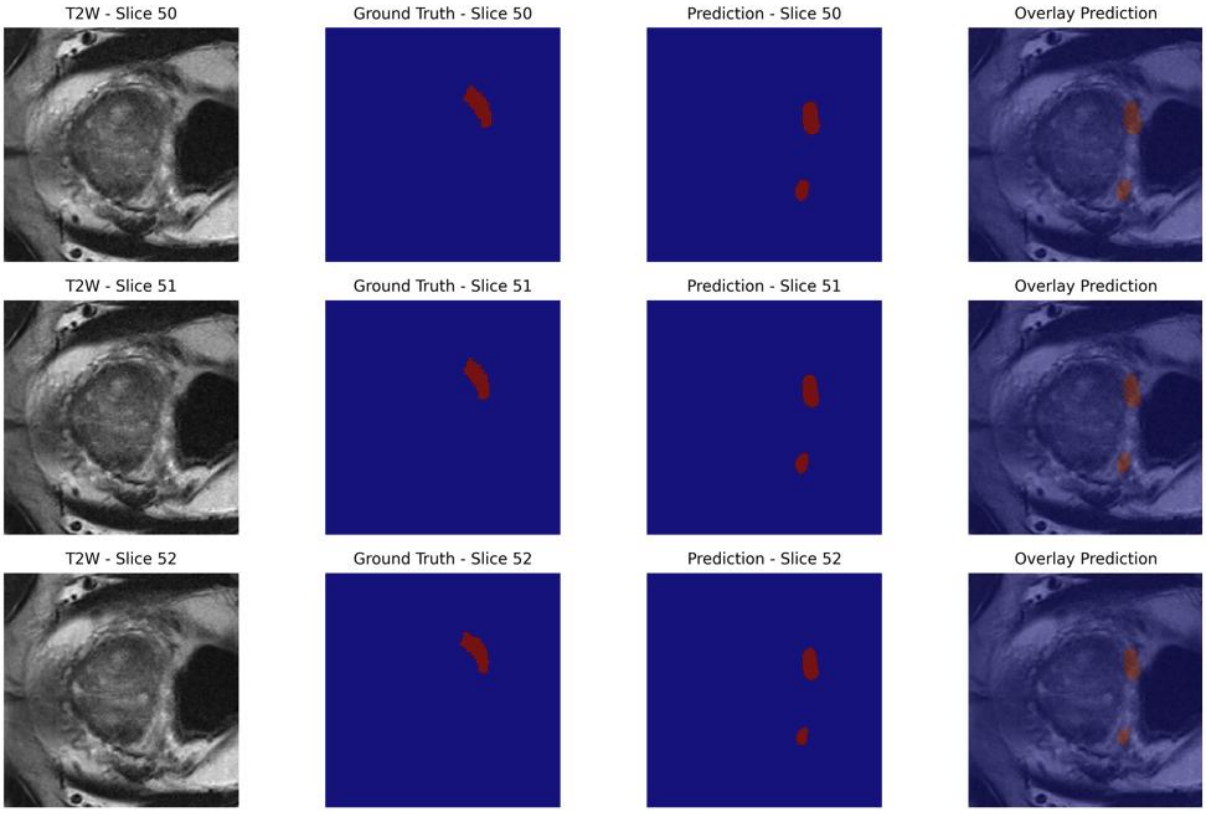

Figure 6 – Visualising a test case where the DSC = 0.4001. The model correctly predicted the existing lesion, but over predicted another lesion.

- Well-Segmented Cases (Dice > 0.5) – see Figure 10 for an example

The top mode of the distribution (~0.5 DSC) was reported for cases where the model achieved strong overlap with the ground truth lesion, especially for mid-gland tumours with evident visual contrast in all mpMRI channels. In these cases, the model accurately identified the lesion's location and border, although with minor over- or under-segmentation on the boundaries.

- Very well segmented cases (Dice > 0.7) – see Figure 9 for an example

Notably, all Dice scores above 0.70 occurred in single-lesion cases, where the U-Net had a single foci to resolve. In these instances, the larger and more distinct peripheral zone tumours were typically segmented with higher overlap and minimal confusion, reflecting the model's relative strength when only one lesion is present.

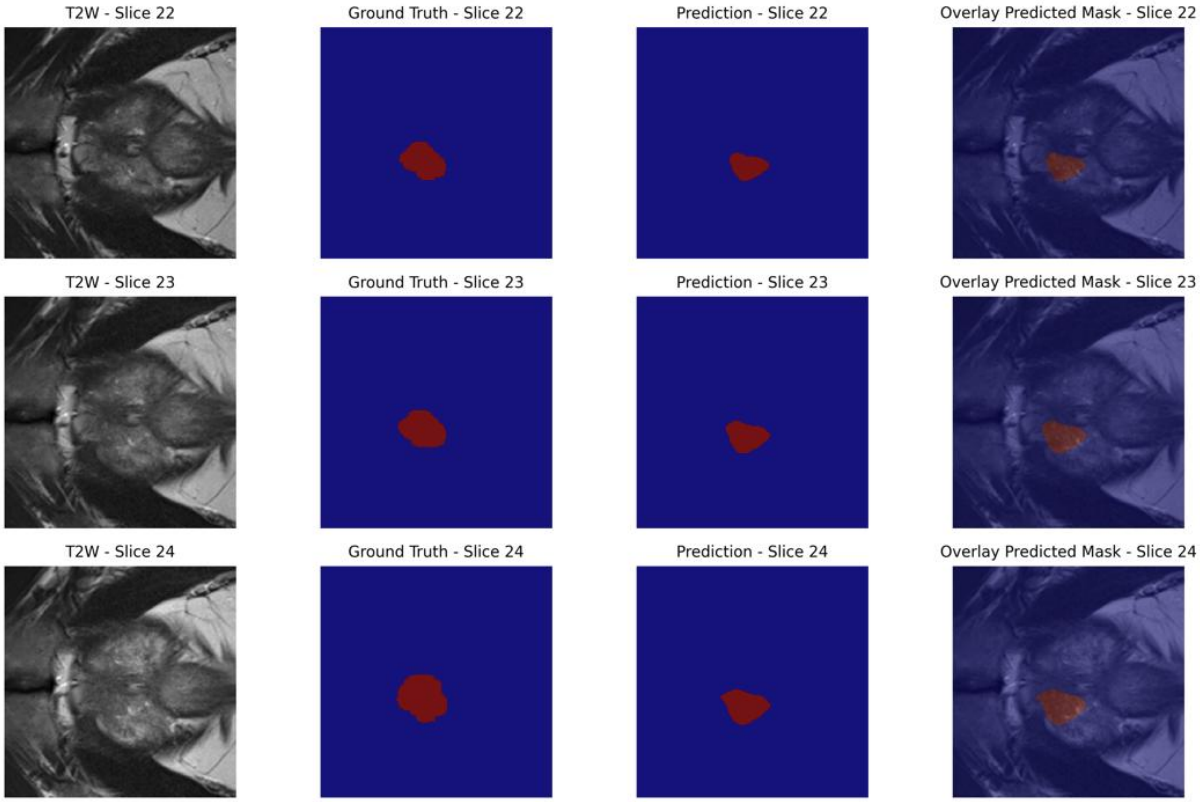

Figure 8 – Visualising a test case where DSC = 0.7682. The model predicted the lesion. The only error is with the precise delineation of the lesion border

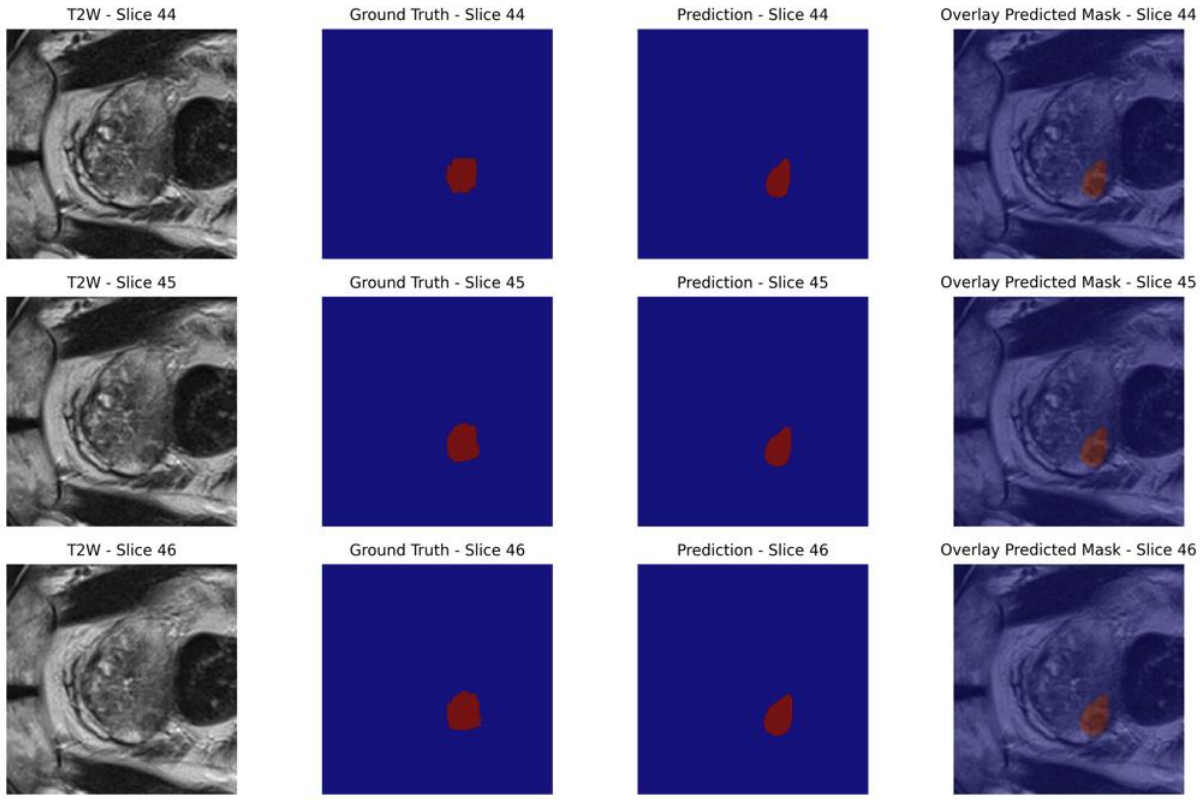

Figure 7 – Visualising a test case were DSC = 0.6397. The model performed broadly well by identifying the lesion, but it is worse at identifying the boundary as well as other examples. This case has the lesion less central than cases with DSC>0.7.

Even with the overall mean DSC of 0.3145 looking low in context, this is strongly skewed by the large number of complete miss cases (DSC = 0). Focusing on the non-zero prediction cases, however, the average DSC holds up well against the inter-rater agreement benchmarks for prostate cancer segmentation.

Moreover, our results highlight the importance of lesion size, location, and visual discriminability in segmentation accuracy prediction. The U-Net model consistently struggled with apex/base tumours and small or subtle lesions but performed reasonably well in detecting large or well-defined lesions in the mid-gland region.

Discussion

Key Findings

The U-Net model's test set Dice similarity coefficients (DSC) on 128 prostate MRI cases were seen to have a distinctly bimodal distribution, with modes around about 0.20 and 0.40–0.50. In practice, this meant the model would either do very poorly (DSC in the ~0.2 range) or have a relatively good overlap (DSC in the ~0.4–0.5 range), with relatively few in between. This bimodal outcome means there were two subgroups of test cases: one where the model failed to segment the lesion correctly, and another where it successfully segmented a reasonable part of the tumour. This is a sign of unstable behaviour due most likely to inherent variation in lesion quality or imaging protocol across patients (Canhoto et al., 2013). In fact, even the upper mode (about 0.45) only shows partial agreement with the ground truth, whereas the lower mode (~0.2) captures near-failure. The results therefore highlight that the model's performance is not uniform across all cases. Below, I comment on potential technical and clinical explanations for this bimodal performance curve. These results directly address Objective 1 by demonstrating the segmentation performance achievable with a simple U-Net model

Tumour size also appears to play a significant role. Small prostate cancer lesions are by their very nature challenging to segment algorithms, and even slight errors can severely reduce the Dice score (Bakas et al., 2018b; Isensee et al., 2021). A small lesion occupies only a few voxels; when the model even partially misses it or mislabels a chunk of overlying tissue, the overlap (DSC) collapses. The low-DSC mode (~0.2) will likely cover most cases with extremely small tumours. In these cases the model's prediction might only partially overlap the true lesion (or miss it altogether), resulting in a low DSC. On the contrary, larger lesions possess more overlap volume; the model is able to capture at least portion of such large tumours and produce higher DSC values (~0.4–0.5). This is in agreement with results that segmentation algorithms produce higher Dice scores for larger tumours than for small tumours (Bakas et al., 2018c; Taha & Hanbury, 2015). Small lesion volume makes the effect of errors on the Dice score greater, resulting in the poor-performing cluster. Clinically, this could be seen as the

model is not accurate at the early-stage or low-disease but is accurate on bulky, obvious tumours.

The majority of patients in the cohort have multiple foci of tumour, which makes the segmentation task more complex (Ahmed et al., 2017). Overlap for all lesions is averaged for a single DSC per patient in such a situation, and if the model does not detect one of multiple lesions, the overall overlap score falls considerably. For instance, in a two-lesion patient, accurately segmenting one of the lesions but not segmenting the other will give a low combined Dice. The model can also combine two adjacent lesions into a single region or divide a lesion into parts, which impacts the measure of overlap (Taha & Hanbury, 2015). Cases of multi-focal disease would be over-represented in the low-DSC mode, as the network could have segmented only a fraction of the lesions (or substantial parts of one lesion but not the other) (Bakas et al., 2018a). This failure to consistently capture all lesions within a case involving multiple lesions would yield Dice scores near the lower modal peak. On the other hand, single large lesion cases are easier – the model is targeting a single object, and a partial correct segmentation of that object yields an intermediate DSC (capturing the greater mode). Lesion number heterogeneity per patient can thus split performance into two subgroups. Note that the Dice score here is computed per patient over all lesions; this conceals instance-wise performance because a single secondary lesion overlooked will subtract from the total score (Reinke et al., 2021). The bimodal distribution would therefore be a mixture of uni-focal cases (moderate DSC) and multi-focal cases (usually low DSC due to at least one missed lesion).

The anatomic zone of the tumour within the prostate is also a potential consideration. Peripheral zone (PZ) lesions will tend to be very conspicuous on diffusion-weighted imaging and ADC maps (and often appear as areas of restricted diffusion), whereas transition zone (TZ) lesions can be subtle and often obscured by benign prostatic hyperplasia nodules (inflamed prostate(Barentsz et al., 2012; Turkbey et al., 2019)). If the training dataset was comprised mostly of PZ tumours (the most common location of prostate cancer), the model might have learned features that more accurately detect PZ lesions. TZ tumours, conversely, might be under-segmented or missed, and would thus have lower Dice scores. Clinical evidence corroborates this discrepancy: even for human readers, there is less sensitivity for detection of TZ tumours on mpMRI than for PZ tumours (Rosenkrantz, Ginocchio, et al., 2016; Vargas et al., 2012). In our results, it is plausible that one of the modes of the Dice distribution can correspond to more observable PZ lesions (giving moderate overlap), and the other mode to TZ or hard-location lesions (giving poor overlap).For instance, an evident posterolateral PZ tumour might be caught partially by the model (DSC ~0.5), whereas an anterior TZ lesion might be lost or just picked up (DSC ~0.1–0.2). Zonal imaging appearance variation and prevalence might then split the performance of the model. This highlights the limitation of the network failing to perform equally across all sub-regions in the anatomy of the prostate.

Additionally, all images were intensity-normalized, co-registered, and resampled, which was unavoidable but would introduce variability or minimal errors in some cases. In the case that MRI sequence co-registration was less than perfect for several patients, the model would receive misaligned multi-parametric data, and therefore segmentation failure (Gholizadeh-Ansari et al., 2020). Quality problems of the image (diffusion-weighted image distortion, motion artifacts) can also detract from model performance. Cases that are poorly performing (low DSC) should likely belong to cases of these artifacts as the model-learned features can degrade on encountering non-standard input data. For example, a case having severe motion blur or noise may mislead the network and lead it to fail to detect the lesion at all. Also, if some patients had very poor tumour-to-background contrast (e.g. a poorly defined lesion on all sequences), the model may not be confident to segment it. In contrast, clean, high-quality imaging and well-behaved preprocessing cases are more often in the higher-DSC group. It is to be kept in mind that the raw data provided was of relatively high quality, with issues such as motion artifacts from patients would be less likely. Errors arising from this reason is less likely to affect the final results.

The bimodal outcome suggests the model is failing/succeeding in two typical manners: in most cases it under-segments or completely fails to segment the lesion (false negatives), and in the remaining ones it outlines the area of the lesion but potentially extends into other non-tumour regions as well (false positives) (Bakas et al., 2018a). In either situation, a low Dice score is generated. For instance, if the network is not highlighting a tumour at all, then the DSC for such an example will be nearly 0. Similarly, if it is segmenting some region which doesn't really belong to tumour (e.g. incorrectly classifying benign tissue as cancer), then the overlap with the actual lesion will be minimal. The cluster at DSC ~0.2 likely includes cases with deep false negatives (the model recognized little to no actual lesion) or mislocalised predictions. The top mode (~0.45) suggests that in those instances the model did recognize the lesion but underestimated its size or included some false-positive regions surrounding it, which resulted in partial overlap. Most notably, the Dice metric is responsive to both types of error: leaving part of the lesion out or misclassifying healthy tissue both reduce the score (Reinke et al., 2021). The sensitivity is greater for small lesions (Bakas et al., 2018a)– a small number of voxels did not make it or are false are sufficient to make the Dice score react strongly.

Collectively, these factors paint a consistent picture: cases of large, well-defined, solitary lesions (often in the peripheral zone) are segmented with decent accuracy, while cases of small, faint, or multiple lesions are segmentation failures (often involving transition zone tumours or poorer image quality). These two peaks in the Dice distribution mirror these two categories of model performance. This justification is supported by earlier studies in intraprostatic lesion segmentation, which reported high performance variability between cases. One of the deep learning algorithms, for example, had a mean DSC of only 0.41 with quite a large standard deviation (±0.28) in prostate tumours (Seetharaman et al., 2021). Showing that some of them could have good overlap while the majority had close to zero – an observation quite consistent with

our bimodal findings. Other studies have achieved DSC 0.35–0.60 for prostate lesion segmentation depending on the cohort and techniques (Isensee et al., 2021; LaBella et al., 2024), again frequently noting that small or invisible tumours are overlooked. Our analysis therefore highlights established failure modes of automated segmentation in this case and offers particular insight into why the model will work on some tumours but not others.

Clinical Relevance

The clinical implications of this work is encouraging. Prostate cancer diagnosis and management increasingly rely on the reading of mpMRI scans by specialists, which is time-consuming and subject to inter-reader variability (Gaur et al., 2021). Our automated segmentation software may be used as a second reader or triage tool, bringing suspicious regions to the attention of the radiologist for review (Schelb et al., 2021).

One potential application is assistance for radiologists in daily MRI reporting. As an example, the AI can pre-segment lesions and provide volume measurements as a starting point for editing by the radiologist, potentially saving time and bringing consistency. Another application is in planning targeted biopsy and treatment. The algorithm output can assist in planning MRI-ultrasound fusion biopsies by demonstrating lesion location and extent (Boldt et al., 2019). Similarly, for focal therapies or radiation dose escalation to dominant lesions, an auto-contour tool could speed up the treatment planning process (with physician approval) (Boldt et al., 2019).

It is important to mention that such an AI program would serve rather than replace the human expert. As it is, the model still misses some cancers and, on occasion, emphasizes benign spots, so a radiologist's analysis is still essential. However, even with these limitations, the tool can improve workflow by quickly providing an initial interpretation. For example, an automated system could flag exams with likely lesions, helping prioritize those for immediate review. By providing consistent segmentations, the AI might also help reduce inter-observer variability – radiologists using the AI's output as a guide could become more consistent with each other in what they consider the lesion's extent (Boldt et al., 2019).

Clinically, the fact that the U-NET model is comparable to prior methods (Bakas et al., 2018c; Seetharaman et al., 2021). The promise shown by this U-Net method shows that even without state-of-the-art techniques, AI is able to extract useful features from mpMRI to aid in cancer detection. As the technology is further refined (e.g., through improvements detailed in the future work section), it can emerge as a reliable part of the diagnostic process. AI segmentations could one day be a standard tool used by radiologists as a "second set of eyes" to double-check their PI-RADS assessments, combining human judgment with AI consistency (Gaur et al., 2021). This synergy can improve overall diagnostic performance – i.e., the AI can pick up on a lesion that a less experienced reader might miss, or, conversely, the radiologist can override an AI false positive lesion detection. In summary, the clinical implication of our findings is that they

bring us one step closer to AI-assisted reading of prostate MRI, which can lead to earlier and more reproducible detection of clinically significant cancers and, ultimately, better patient outcomes (Boldt et al., 2019). This reinforces Objective 3, suggesting AI tools like ours could meaningfully enhance efficiency in clinical prostate cancer workflows.

Limitations and Future Work

The model's architecture may be enhanced. Our U-Net was simple; the incorporation of current breakthroughs may address performance deficits. Utilizing attention mechanisms (such as Attention U-Net) might enable attention to the right regions and boundaries, for instance (Oktay et al., 2018c). Transformer architectures for segmentation (currently emerging in research) might capture long-range dependencies and enhance separation of complex shapes. Research into such architectures for prostate MRI is a natural extension. Likewise, experimenting with loss functions is worthwhile. One could also experiment with boundary-aware loss functions (e.g., variants of Dice or Tversky that penalize boundary errors) to tackle the contour precision issue (Salehi et al., 2017).

Handling small lesions is a clear area for improvement. Data augmentation or specialized training strategies might help the model recognize very small lesions (e.g., oversampling slices with tiny lesions, or using multi-task learning to simultaneously detect lesions) . There is another idea that is a separate detection module: classify with a classification CNN to label slices or regions likely having a lesion, then segment. Multi-task learning (having the network predict also whether there is a lesion on a slice or a PI-RADS category) may force it to learn more robust features for detection (Bakas et al., 2018a). From an implementation standpoint, integration of the model into a clinical workflow and obtaining user feedback will be critical.

Another important step toward future research is the inclusion of healthy patient data in training. In the majority of clinical settings, a high percentage of prostate MRI scans prove to be normal or showing benign conditions (Litjens et al., 2017). By including these non-cancer cases in the model's training process, the network would learn to recognize normal anatomy and benign variants more precisely, which could further reduce false positives. This would better reflect actual clinical distributions, in which only a limited number of patients have suspicious lesions and some scans have no cancer at all (Gaur et al., 2021). As a result, training on healthy patient data can enhance the system's robustness and specificity so that it better separates actual cancers from benign findings without sacrificing stable sensitivity for clinically relevant disease.

Having a prototype where radiologists can see AI-generated contours overlaid on MRI and provide comments would guide the refinements (Azimi & Pahl, 2020). Hypothetically, radiologists might prefer to have a confidence level with each prediction or visualizations detailing (heatmaps) why the model highlighted a region (Samek et al., 2017).

To assess the clinical utility of the proposed segmentation model, future studies should look beyond technical performance metrics and investigate its impact on the diagnostic process. One such potential area is developing reader studies in which trained radiologists read prostate MRI cases with and without the assistance of AI-guided segmentations (Ungi et al., 2020). Such studies would allow systematic evaluation of whether AI support might improve cancer detection rates, reduce reading time, or boost confidence in clinical decision-making. It would also allow investigation of any cognitive bias that would be introduced by the AI, e.g., over-reliance on the model's suggestion (Hekler et al., 2019). These evaluations are important in determining not only how well the model works in isolation, but how well it integrates into real-world clinical practice and whether it has a measurable positive effect on patient care.

Lastly, the limitations of our study (data set, 2D architecture, straightforward model design) are clear avenues for future growth. By expanding the data set, embracing 3D/complicated models, focusing on small lesion detection, and integrating the tool into the workflow (with proper evaluation), we can increase the performance as well as usability of the system . The developments attained in this project lay a solid groundwork. Implementing these future enhancements, we move closer to a stronger AI system capable of assisting clinicians more effectively in prostate cancer detection on MRI.

Conclusion

This research project demonstrates that deep learning can be effectively applied to prostate mpMRI for lesion segmentation, even using a relatively simple U-Net model. The CNN-based U-Net segmentation model established that it can identify and segment prostate lesions with moderate accuracy (averaging around 30% overlap with expert radiologist defined tumour regions). Whilst the performance of this U-Net performance is not perfect by any means, it is a positive outcome given the complexity of prostate lesions and comparable to levels of agreement between radiologists. Thus, the first objective – to test the accuracy of the model in a simple U-Net – was successful with encouraging findings: the model is an effective tool in the detection of suspicious areas on MRI scans, and can act as a guide to accompany radiologists opinions.

We found that even this relatively simple U-Net achieved moderate Dice scores (~0.31 on average), aligning with inter-observer agreement levels. Therefore, one of the main conclusions we can draw from the U-Net's performance with mpMRI prostate images, is it represents a solid trade-off between accuracy and efficiency, with its ability to identify the existence of lesions. Training the model took roughly 20 hours on a single GPU and the inference is relatively instant – the model can be integrated into existing workflows with very little delay. Furthermore, despite its simplicity, the model achieved a solid accuracy, thus it serves as a strong starting point for an AI assisted segmentation tool for radiologists. However, as discussed, the model struggles with smaller or multi-focal lesions and may require advanced architectures or training strategies to address these shortcomings. These findings affirm the promise of applying AI to prostate MRI, echoing

prior literature that deep learning can improve detection and segmentation of prostate lesions. Because of its quick inference time (~instant segmentation once trained), the model could feasibly be embedded in routine practice to highlight suspicious regions. Future work must quantify actual time saved, biopsy reduction, and radiologist confidence gains to confirm an overall efficiency improvement.

Statement of Student Contribution

My original contributions include but are not limited to:

- Pre-processing the mpMRI dataset
- Designing and implementing the U-Net model from scratch
- Training the U-Net model
- Evaluating the segmentation accuracy using the key metrics
- I optimised the U-Net model to improve model performance and implemented visualisation techniques for comparing the segmentation outputs

My work ensured that the U-Net model was accurate and computationally efficient.

References

I acknowledge the use of ChatGPT(version GPT-4o, Microsoft, https://chatgpt.com/) to summarise my initial notes and to proofread my final draft.

Ahmed, H. U., El-Shater Bosaily, A., Brown, L. C., Gabe, R., Kaplan, R., Parmar, M. K., Collaco-Moraes, Y., Ward, K., Hindley, R. G., Freeman, A., Kirkham, A. P., Oldroyd, R., Parker, C., & Emberton, M. (2017). Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *The Lancet*, *389*, 815–822. https://doi.org/10.1016/S0140-6736(16)32401-1

Alqahtani, S. (2024). Systematic Review of AI-Assisted MRI in Prostate Cancer Diagnosis: Enhancing Accuracy Through Second Opinion Tools. *Diagnostics (Basel, Switzerland)*, *14*. https://doi.org/10.3390/diagnostics14222576

Azimi, S., & Pahl, C. (2020). A layered quality framework for machine learning-driven data and information models. *ICEIS 2020 - Proceedings of the 22nd International Conference on Enterprise Information Systems*, *1*, 579–587. https://doi.org/10.5220/0009472305790587

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M., Prastawa, M., Alberts, E., Lipkova, J., Freymann, J., Kirby, J., Bilello, M., Fathallah-Shaykh, H., Wiest, R., Kirschke, J., … Menze, B. (2018a). *Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge*.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M., Prastawa, M., Alberts, E., Lipkova, J., Freymann, J., Kirby, J., Bilello, M., Fathallah-Shaykh, H., Wiest, R., Kirschke, J., … Menze, B. (2018b). *Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge*.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M., Prastawa, M., Alberts, E., Lipkova, J., Freymann, J., Kirby, J., Bilello, M., Fathallah-Shaykh, H., Wiest, R., Kirschke, J., … Menze, B. (2018c). *Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge*.

Barentsz, J. O., Richenberg, J., Clements, R., Choyke, P., Verma, S., Villeirs, G., Rouviere, O., Logager, V., & Fütterer, J. J. (2012). ESUR prostate MR guidelines 2012. *European Radiology*, *22*, 746–757. https://doi.org/10.1007/s00330-011-2377-y

Bergengren, O., Pekala, K. R., Matsoukas, K., Fainberg, J., Mungovan, S. F., Bratt, O., Bray, F., Brawley, O., Luckenbaugh, A. N., Mucci, L., Morgan, T. M., & Carlsson, S. V. (2023). 2022 Update on Prostate Cancer Epidemiology and Risk Factors—A Systematic Review. In *European Urology* (Vol. 84, pp. 191–206). Elsevier B.V. https://doi.org/10.1016/j.eururo.2023.04.021

Bhayana, R., O'Shea, A., Anderson, M. A., Bradley, W. R., Gottumukkala, R. V., Mojtahed, A., Pierce, T. T., & Harisinghani, M. (2021). PI-RADS versions 2 and 2.1: Interobserver agreement and diagnostic performance in peripheral and transition zone lesions among six radiologists. *American Journal of Roentgenology*, *217*, 141–151. https://doi.org/10.2214/AJR.20.24199

Boldt, A., Schiffer, A. M., Waszak, F., & Yeung, N. (2019). Confidence Predictions Affect Performance Confidence and Neural Preparation in Perceptual Decision Making. *Scientific Reports*, *9*. https://doi.org/10.1038/s41598-019-40681-9

Cancer in Men: Prostate Cancer is #1 for 118 Countries Globally. (n.d.). Https://Www.Cancer.Org/Research/Acs-Research-News/Prostate-Cancer-Is-Number-1-for-118-Countries-Worldwide.Html.

Canhoto, A. I., Clark, M., & Fennemore, P. (2013). Emerging segmentation practices in the age of the social customer. *Journal of Strategic Marketing*, *21*, 413–428. https://doi.org/10.1080/0965254X.2013.801609

Cao, R., Zhong, X., Afshari, S., Felker, E., Suvannarerg, V., Tubtawee, T., Vangala, S., Scalzo, F., Raman, S., & Sung, K. (2021). Performance of Deep Learning and Genitourinary Radiologists in Detection of Prostate Cancer Using 3-T Multiparametric Magnetic Resonance Imaging. *Journal of Magnetic Resonance Imaging*, *54*, 474–483. https://doi.org/10.1002/jmri.27595

Chareshneu, A., Cantara, A., Tichý, D., & Sehnal, D. (2024). Visualizing Volumetric and Segmentation Data using Mol* Volumes & Segmentations 2.0. *Current Protocols*, *4*. https://doi.org/10.1002/cpz1.70070

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision Transformer: Reinforcement Learning via Sequence Modeling. *Advances in Neural Information Processing Systems*, *18*, 15084–15097.

Chen, M. Y., Woodruff, M. A., Dasgupta, P., & Rukin, N. J. (2020). Variability in accuracy of prostate cancer segmentation among radiologists, urologists, and scientists. *Cancer Medicine*, *9*, 7172–7182. https://doi.org/10.1002/cam4.3386

David MK, & Leslie SW. (2024). 8. prostateNBK557495. *Prostate Specific Antigen*. .

Dickinson, H., Ham, C., Snelling, I., & Spurgeon, P. (2013a). *Are We There Yet? Models of Medical Leadership and their effectiveness: An Exploratory Study*.

Dickinson, H., Ham, C., Snelling, I., & Spurgeon, P. (2013b). *Are We There Yet? Models of Medical Leadership and their effectiveness: An Exploratory Study*.

Dickinson, L., Ahmed, H. U., Kirkham, A. P., Allen, C., Freeman, A., Barber, J., Hindley, R. G., Leslie, T., Ogden, C., Persad, R., Winkler, M. H., & Emberton, M. (2013). A multi-centre prospective development study evaluating focal therapy using high intensity focused ultrasound for localised prostate cancer: The INDEX study. *Contemporary Clinical Trials*, *36*, 68–80. https://doi.org/10.1016/j.cct.2013.06.005

El-Shater Bosaily, A., Parker, C., Brown, L. C., Gabe, R., Hindley, R. G., Kaplan, R., Emberton, M., Ahmed, H. U., Emberton, M., Ahmed, H., Bosaily, A. E. S., Kirkham, A., Freeman, A., Jameson, C., Hindley, R., Parker, C., Cooper, C., Oldroyd, R., Kaplan, R., … Agarwal, S. (2015). PROMIS - Prostate MR imaging study: A paired validating cohort study evaluating the role of multi-parametric MRI in men with clinical suspicion of prostate cancer. *Contemporary Clinical Trials*, *42*, 26–40. https://doi.org/10.1016/j.cct.2015.02.008

Fassia, M.-K., Balasubramanian, A., Woo, S., Vargas, H. A., Hricak, H., Konukoglu, E., & Becker, A. S. (2024). Deep Learning Prostate MRI Segmentation Accuracy and Robustness: A Systematic Review. *Radiology. Artificial Intelligence*, *6*, e230138. https://doi.org/10.1148/ryai.230138

Gaur, L., Afaq, A., Singh, G., & Dwivedi, Y. K. (2021). Role of artificial intelligence and robotics to foster the touchless travel during a pandemic: a review and research agenda. *International Journal of Contemporary Hospitality Management*, *33*, 4079–4098. https://doi.org/10.1108/IJCHM-11-2020-1246

Gholizadeh-Ansari, M., Alirezaie, J., & Babyn, P. (2020). Deep Learning for Low-Dose CT Denoising Using Perceptual Loss and Edge Detection Layer. *Journal of Digital Imaging*, *33*, 504–515. https://doi.org/10.1007/s10278-019-00274-4

GitHub - janishar/mit-deep-learning-book-pdf: MIT Deep Learning Book in PDF format (complete and parts) by Ian Goodfellow, Yoshua Bengio and Aaron Courville. (n.d.). Https://Github.Com/Janishar/Mit-Deep-Learning-Book-Pdf.

Glorot, X., & Bengio, Y. (n.d.). *Understanding the difficulty of training deep feedforward neural networks*. http://www.iro.umontreal.

Group, T. 9 H. T. for M. | N. C. F. M., & urgeinteractive. (n.d.). *Prostate Cancer | North Coast Family Medical Group*. Https://Www.Ncfmg.Com/Top-9-Health-Tips-for-Men/.

Hekler, A., Utikal, J. S., Enk, A. H., Solass, W., Schmitt, M., Klode, J., Schadendorf, D., Sondermann, W., Franklin, C., Bestvater, F., Flaig, M. J., Krahl, D., von Kalle, C., Fröhling, S., & Brinker, T. J. (2019). Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer*, *118*, 91–96. https://doi.org/10.1016/j.ejca.2019.06.012

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, *18*, 203–211. https://doi.org/10.1038/s41592-020-01008-z

Kasivisvanathan, V., Rannikko, A. S., Borghi, M., Panebianco, V., Mynderse, L. A., Vaarala, M. H., Briganti, A., Budäus, L., Hellawell, G., Hindley, R. G., Roobol, M. J., Eggener, S., Ghei, M., Villers, A., Bladou, F., Villeirs, G. M., Virdi, J., Boxler, S., Robert, G., … Moore, C. M. (2018). MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis. *New England Journal of Medicine*, *378*, 1767–1777. https://doi.org/10.1056/nejmoa1801993

Kayalibay, B., Jensen, G., & van der Smagt, P. (2017). *CNN-based Segmentation of Medical Imaging Data*.

Key Statistics for Prostate Cancer | Prostate Cancer Facts. (n.d.). Https://Www.Cancer.Org/Cancer/Types/Prostate-Cancer/about/Key-Statistics.Html.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Kokkinos, V., Sisterson, N. D., Wozny, T. A., & Richardson, R. M. (2019). Association of Closed-Loop Brain Stimulation Neurophysiological Features with Seizure Control among Patients with Focal Epilepsy. *JAMA Neurology*, *76*, 800–808. https://doi.org/10.1001/jamaneurol.2019.0658

LaBella, D., Baid, U., Khanna, O., McBurney-Lin, S., McLean, R., Nedelec, P., Rashid, A., Tahon, N. H., Altes, T., Bhalerao, R., Dhemesh, Y., Godfrey, D., Hilal, F., Floyd, S., Janas, A., Kazerooni, A. F., Kirkpatrick, J., Kent, C., Kofler, F., … Calabrese, E. (2024). *Analysis of the BraTS 2023 Intracranial Meningioma Segmentation Challenge*. https://doi.org/10.59275/j.melba.2025-bea1

Lin, K. H., Hsieh, T. Y., Chen, C. H., & Pu, Y. S. (2023). Digital Rectal Examination Still Plays a Crucial Role of Predicting Outcomes in the Prostate Cancer Patients Undergoing Primary Total Prostate Cryoablation. *Urological Science*, *34*, 187–193. https://doi.org/10.4103/UROS.UROS_139_22

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. In *Medical Image Analysis* (Vol. 42, pp. 60–88). Elsevier B.V. https://doi.org/10.1016/j.media.2017.07.005

Lucido, J. J., DeWees, T. A., Leavitt, T. R., Anand, A., Beltran, C. J., Brooke, M. D., Buroker, J. R., Foote, R. L., Foss, O. R., Gleason, A. M., Hodge, T. L., Hughes, C. O., Hunzeker, A. E., Laack, N. N., Lenz, T. K., Livne, M., Morigami, M., Moseley, D. J., Undahl, L. M., … Patel, S. H. (2023). Validation of clinical acceptability of deep-learning-based automated segmentation of organs-at-risk for head-and-neck radiotherapy treatment planning. *Frontiers in Oncology*, *13*. https://doi.org/10.3389/fonc.2023.1137803

Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, 565–571. https://doi.org/10.1109/3DV.2016.79

Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., Fox, N. C., & Ourselin, S. (2010). Fast free-form deformation using graphics processing units. *Computer Methods and Programs in Biomedicine*, *98*, 278–284. https://doi.org/10.1016/j.cmpb.2009.09.002

Montazerolghaem, M., Sun, Y., Sasso, G., & Haworth, A. (2023). U-Net Architecture for Prostate Segmentation: The Impact of Loss Function on System Performance. *Bioengineering*, *10*. https://doi.org/10.3390/bioengineering10040412

Oktay, O., Schlemper, J., Folgoc, L. Le, Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018a). *Attention U-Net: Learning Where to Look for the Pancreas*.

Oktay, O., Schlemper, J., Folgoc, L. Le, Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018b). *Attention U-Net: Learning Where to Look for the Pancreas*.

Oktay, O., Schlemper, J., Folgoc, L. Le, Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018c). *Attention U-Net: Learning Where to Look for the Pancreas*.

Orczyk, C., Barratt, D., Brew-Graves, C., Peng Hu, Y., Freeman, A., McCartan, N., Potyka, I., Ramachandran, N., Rodell, R., Williams, N. R., Emberton, M., & Ahmed, H. U. (2021). Prostate Radiofrequency Focal Ablation (ProRAFT) Trial: A Prospective Development Study Evaluating a Bipolar Radiofrequency Device to Treat Prostate Cancer. *Journal of Urology*, *205*, 1090–1099. https://doi.org/10.1097/JU.0000000000001567

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Penzkofer, T. (2024). Prostate-MRI reporting should be done with the aid of AI systems: Pros. *European Radiology*, *34*, 7728–7730. https://doi.org/10.1007/s00330-024-10909-y

PROMIS Study Dataset - Open Access Request - NCITA. (n.d.). Https://Ncita.Org.Uk/Promis-Data-Set-Open-Access-Request/.

Prostate MRI - RAD-ASSIST. (n.d.). Https://Rad.Bwh.Harvard.Edu/Prostate-Mri/.

Reinke, A., Tizabi, M. D., Sudre, C. H., Eisenmann, M., Rädsch, T., Baumgartner, M., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Bankhead, P., Benis, A., Blaschko, M., Buettner, F., Cardoso, M. J., Chen, J., Cheplygina, V., Christodoulou, E., Cimini, B., … Maier-Hein, L. (2021). *Common Limitations of Image Processing Metrics: A Picture Story*.

Ronneberger, O., Fischer, P., & Brox, T. (2021). INet: Convolutional Networks for Biomedical Image Segmentation. *IEEE Access*, *9*, 16591–16603. https://doi.org/10.1109/ACCESS.2021.3053408

Rosenkrantz, A. B., Ginocchio, L. A., Cornfeld, D., Froemming, A. T., Gupta, R. T., Turkbey, B., Westphalen, A. C., Babb, J. S., & Margolis, D. J. (2016). Interobserver reproducibility of the PI-RADS version 2 lexicon: A multicenter study of six experienced prostate radiologists. *Radiology*, *280*, 793–804. https://doi.org/10.1148/radiol.2016152542

Rosenkrantz, A. B., Verma, S., Choyke, P., Eberhardt, S. C., Eggener, S. E., Gaitonde, K., Haider, M. A., Margolis, D. J., Marks, L. S., Pinto, P., Sonn, G. A., & Taneja, S. S. (2016). Prostate Magnetic Resonance Imaging and Magnetic Resonance Imaging Targeted Biopsy in

Patients with a Prior Negative Biopsy: A Consensus Statement by AUA and SAR. *Journal of Urology*, *196*, 1613–1618. https://doi.org/10.1016/j.juro.2016.06.079

Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). Tversky loss function for image segmentation using 3D fully convolutional deep networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10541 LNCS*, 379–387. https://doi.org/10.1007/978-3-319-67389-9_44

Samek, W., Wiegand, T., & Müller, K.-R. (2017). *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*.

Sarkar, K., & Li, B. (2022). Deep learning for medical image segmentation. In *Research Anthology on Improving Medical Imaging Techniques for Analysis and Intervention* (pp. 861–891). IGI Global. https://doi.org/10.4018/978-1-6684-7544-7.ch044

Schelb, P., Tavakoli, A. A., Tubtawee, T., Hielscher, T., Radtke, J. P., Görtz, M., Schütz, V., Kuder, T. A., Schimmöller, L., Stenzinger, A., Hohenfellner, M., Schlemmer, H. P., & Bonekamp, D. (2021). Comparison of prostate MRI lesion segmentation agreement between multiple radiologists and a fully automatic deep learning system. *RoFo Fortschritte Auf Dem Gebiet Der Rontgenstrahlen Und Der Bildgebenden Verfahren*, *193*, 559–573. https://doi.org/10.1055/a-1290-8070

Seetharaman, A., Bhattacharya, I., Chen, L. C., Kunder, C. A., Shao, W., Soerensen, S. J. C., Wang, J. B., Teslovich, N. C., Fan, R. E., Ghanouni, P., Brooks, J. D., Too, K. J., Sonn, G. A., & Rusu, M. (2021). Automated detection of aggressive and indolent prostate cancer on magnetic resonance imaging. *Medical Physics*, *48*, 2960–2972. https://doi.org/10.1002/mp.14855

Simmons, L. A. M., Kanthabalan, A., Arya, M., Briggs, T., Barratt, D., Charman, S. C., Freeman, A., Hawkes, D., Hu, Y., Jameson, C., McCartan, N., Moore, C. M., Punwani, S., van der Muelen, J., Emberton, M., & Ahmed, H. U. (2018). Accuracy of Transperineal Targeted Prostate Biopsies, Visual Estimation and Image Fusion in Men Needing Repeat Biopsy in the PICTURE Trial. *Journal of Urology*, *200*, 1227–1234. https://doi.org/10.1016/j.juro.2018.07.001

Sotiras, A., Davatzikos, C., & Paragios, N. (2013). Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, *32*, 1153–1190. https://doi.org/10.1109/TMI.2013.2265603

Sun, Z., Wang, K., Kong, Z., Xing, Z., Chen, Y., Luo, N., Yu, Y., Song, B., Wu, P., Wang, X., Zhang, X., & Wang, X. (2023a). A multicenter study of artificial intelligence-aided software for detecting visible clinically significant prostate cancer on mpMRI. *Insights into Imaging*, *14*. https://doi.org/10.1186/s13244-023-01421-w

Sun, Z., Wang, K., Kong, Z., Xing, Z., Chen, Y., Luo, N., Yu, Y., Song, B., Wu, P., Wang, X., Zhang, X., & Wang, X. (2023b). A multicenter study of artificial intelligence-aided software for detecting visible clinically significant prostate cancer on mpMRI. *Insights into Imaging*, *14*. https://doi.org/10.1186/s13244-023-01421-w

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, *15*. https://doi.org/10.1186/s12880-015-0068-x

Turkbey, B., Rosenkrantz, A. B., Haider, M. A., Padhani, A. R., Villeirs, G., Macura, K. J., Tempany, C. M., Choyke, P. L., Cornud, F., Margolis, D. J., Thoeny, H. C., Verma, S., Barentsz, J., & Weinreb, J. C. (2019). Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. In *European Urology* (Vol. 76, pp. 340–351). Elsevier B.V. https://doi.org/10.1016/j.eururo.2019.02.033

Twilt, J. J., van Leeuwen, K. G., Huisman, H. J., Fütterer, J. J., & de Rooij, M. (2021). Artificial Intelligence Based Algorithms for Prostate Cancer Classification and Detection on Magnetic Resonance Imaging: A Narrative Review. *Diagnostics (Basel, Switzerland)*, *11*. https://doi.org/10.3390/diagnostics11060959

Ungi, T., Greer, H., Sunderland, K. R., Wu, V., Baum, Z. M. C., Schlenger, C., Oetgen, M., Cleary, K., Aylward, S. R., & Fichtinger, G. (2020). Automatic Spine Ultrasound Segmentation for Scoliosis Visualization and Measurement. *IEEE Transactions on Biomedical Engineering*, *67*, 3234–3241. https://doi.org/10.1109/TBME.2020.2980540

Vargas, H. A., Akin, O., Franiel, T., Goldman, D. A., Udo, K., Touijer, K. A., Reuter, V. E., & Hricak, H. (2012). Normal central zone of the prostate and central zone involvement by prostate cancer: Clinical and mr imaging implications. *Radiology*, *262*, 894–902. https://doi.org/10.1148/radiol.11110663

Villanueva-Meyer, J. E., Mabray, M. C., & Cha, S. (2017). Current clinical brain tumor imaging. *Clinical Neurosurgery*, *81*, 397–415. https://doi.org/10.1093/neuros/nyx103

Weinreb, R. N., Leung, C. K. S., Crowston, J. G., Medeiros, F. A., Friedman, D. S., Wiggs, J. L., & Martin, K. R. (2016). Primary open-angle glaucoma. *Nature Reviews. Disease Primers*, *2*, 16067. https://doi.org/10.1038/nrdp.2016.67

Yan, W., Chiu, B., Shen, Z., Yang, Q., Syer, T., Min, Z., Punwani, S., Emberton, M., Atkinson, D., Barratt, D. C., & Hu, Y. (2023). *Combiner and HyperCombiner Networks: Rules to Combine Multimodality MR Images for Prostate Cancer Localisation*. https://doi.org/10.1016/j.media.2023.103030

Yousef, R., Khan, S., Gupta, G., Siddiqui, T., Albahlal, B. M., Alajlan, S. A., & Haq, M. A. (2023). U-Net-Based Models towards Optimal MR Brain Image Segmentation. *Diagnostics (Basel, Switzerland)*, *13*. https://doi.org/10.3390/diagnostics13091624

Zhang, H., Patkar, S., Lis, R., Merino, M. J., Pinto, P. A., Choyke, P. L., Turkbey, B., & Harmon, S. (2024). Masked Image Modeling Meets Self-Distillation: A Transformer-Based Prostate Gland Segmentation Framework for Pathology Slides. *Cancers*, *16*. https://doi.org/10.3390/cancers16233897

Acknowledgements

I would like to express my deepest appreciation to Yipei Wang, for without her help I would be very lost. This research project would not have been completed if not for her guidance and inputs.

I appreciate Yipeng Hu also for being my personal tutor, providing me extra-curricular support throughout my time at UCL.

I am also grateful to my classmates, for their moral support.

Appendix

All the code is available on GitHub:

https://github.com/parthiv-n/Research-Project/tree/main

The patient data cannot be shared publicly