QATAR UNIVERSITY

COLLEGE OF ENGINEERING

A VISUALIZATION TRAINING TOOL FOR OPTIMUM CUTTING RANGE DURING URETHRAL DISSECTION IN ROBOT-ASSISTED RADICAL PROSTATECTOMY: A DEEP LEARNING APPROACH

BY

SARAH LOTFI KHARBACH

A Thesis Submitted to

the College of Engineering

in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Computing

June 2021

COMMITTEE PAGE

The members of the Committee approve the Thesis of
Sarah Lotfi Kharbach defended on 18/04/2021.

Dr. Abdulla Al-Ali
Thesis Supervisor

Dr. Uvais Qidwai
Committee Member

Dr. Dena Al-Thani
Committee Member

Approved:

Khalid Kamal Naji, Dean, College of Engineering

ABSTRACT

Kharbach, Sarah, L., Masters : June: 2021, Master of Science in Computing

Title: A Visualization Training Tool for Optimum Cutting Range During Urethral Dissection in Robot-Assisted Radical Prostatectomy: A deep learning approach

Supervisor of Thesis: Dr. Abdulla Al-Ali.

Surgical training and skills assessment is one of the methodologies used by clinical practitioners to master their skills in a fault-tolerant, safe and risk-free environment. Software training tools is one of them. Narrowing the scope on surgery, and with the significant popularity of robotic surgeries in the last few years, clinical and technological research has shifted their attention towards developing surgical supportive tools such intraoperative surgical planning, training tools, etc. In this thesis, we attempt to develop a software to teach surgeons on the optimum urethra dissection locations based on prostate size and cancer location. The surgeon is presented by a patient case and suggests optimum different dissection locations. The system would automatically evaluate the surgeon's performance and generate a score accordingly. In this thesis, a clinical validation was conducted to validate the need of such software in a clinical practice by interviewing three robotic surgeons experts. Once validated successfully, Tutorial Module was developed and validated with expert urology surgeons using content validity to evaluate the module's effectiveness as a teaching modality and Questionnaire for User Interaction Satisfaction to evaluate the interface perceptiveness among them. Both tests reflected high surgeons' satisfaction. To detect the urethra, U-Net, a deep learning module was trained over a small dataset of images extracted from from HMC's operating rooms as well as publicly available videos. Image semantic segmentation was the method used to extract and locate the urethra. U-Net model was able to reach a Jaccard Coefficient of 78.42% and a Dice Coefficient of 88.11%. To the best of our knowledge, this thesis is the first of its kind to offer a scenario-based training on optimum urethra dissection location for robot-assisted radical proastatectomy.

DEDICATION

to my dear beautiful parents, this one is for you.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude and thankfulness to Allah, the Mighty, for guiding me through the right path and helping me endlessly throughout this journey.

Additionally, it is with great gratitude and pleasure that I reserve these few lines as appreciation to all those who have helped me throughout my thesis: Dr. Abdulla for being the major support and patience throughout different thesis phases including topic changes; for his guidance, insightful feedback, promptness and unlimited support in all aspects. Thesis would not be possible without him.

I would like to thank my parents and my 3 sisters: Mariam, Asma and Fatouma for their unlimited patience and support during this journey. I would like to extend my thanks to my **very, very** supportive friends round me and for their continuous help and patience; especially Ealaf Hussein, Salma Shalaby, Mona Youssef, Daniel Zeitouny, etc. I cannot go without thanking my supportive work colleagues for supporting my decisions, guiding me though and facilitating the hardships for me to successfully finish the project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Medical skills training and assessment have become an integral part of medical education in the last decades. Such tools allow clinicians to continuously improve their clinical skills and stay of the edge of latest trends in their fields. Medical training comes in numerous shapes: training can be on cadavers, phantoms, typical lectures and exams, online courses, high-fidelity simulations such as Virtual Reality (VR) and Augmented Reality (AR), computer software, etc. Taking robotic surgeries for instance, training for such procedures is gaining unprecedented attention due to the benefits of robotic surgeries. This is because of the rising adaptation of robotic interventions in many surgical fields such as urology and gynecology. For instance, and with COVID-19 global pandemic affecting the world and specifically the healthcare system, robotic surgeries were an optimum solution to physically separate the healthcare worker and the patient under operation and mitigate infectious contamination as much as possible [1].

Prostatectomy is a medical intervention where the surgeon surgically removes a part or the whole cancerous prostate gland. This procedure can be done robotically, laparoscopically or even as an open surgery. Robotically, Robot-Assisted Radical Prostatectomy or RARP procedure has several steps including releasing the bladder, endopelvic fascia, urethral transection or dissecting [2]. Urethral dissection is one of the most crucial step of the procedure as failing to cut the urethra on the correct location will compromise the surgery outcomes: leaving cancerous tissues behind and failure of the radical extraction of cancer. In this thesis, we are attempting to build a training tool for surgeons to teach them the optimum cutting ranges on the urethra based on two variables: prostate size and cancer location in the prostate. The project has two major components: Tutorial mode and Evaluation mode. Tutorial mode will serve as a tool to teach surgeons about the cutting decisions. Evaluation mode will assess surgeons' acquired knowledge by allowing them to suggest a cutting location and assigning them a score. Deep learning is used to automate the evaluation, as the urethra image will be fed to a trained U-Net model, segment the urethra, draw a box around it, and prompt the trainee surgeon to suggest an excellent, acceptable, unacceptable cut and the urethral stump. By the end of this training, the trainee surgeon is expected to know the optimum cutting location based on each patient case.

The objectives of this thesis are as follows:

1. to provide a solution for trainee surgeons to teach the fundamentals of optimum cutting ranges. The cutting decisions are based on two essential factors: prostate size and cancer location in the prostate. These factors were chosen after iterative discussions with an expert RARP surgeon in Hamad Medical Corporation (HMC).

2. to investigate automated guidance and training approaches using deep learning while leveraging the existing resources of RARP's surgical videos for continuous training and development purposes. This would be responsible of detecting the urethra from a series of images to familiarize trainee surgeons with pelvic area anatomy, specifically urethra.

3. to validate the prototype by conducting evaluation studies for it to be ready for deployment in HMC.

The reminder of the document is structured as follows: Chapter 2 will have clinical and technical backgrounds on most of the concepts mentioned later in thesis. Chapter 3 will have a literature review on state-of-art in training in light of robotic surgery and in medical image segmentation field. Chapter 4 proceeds to discuss the urethra localisation methodology using a deep learning approach and reports key findings. Chapter 5 discusses the clinical validation process that has been conducted before and during the development. Lastly, chapter 6 concludes the key findings, lists the limitations, foresee future work and anticipated publications.

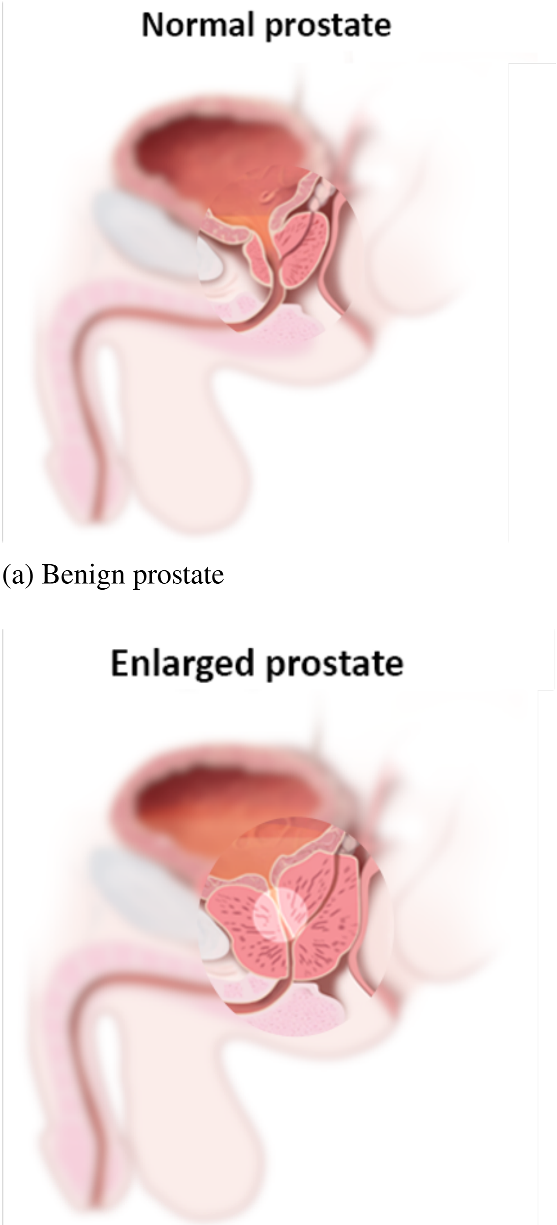
CHAPTER 2: BACKGROUND

This chapter discuss key clinical and technical backgrounds needed to understand thesis topics discussed in later chapters.

2.1 Clinical Background

The section elaborates on the clinical background of the problem in study. Thesis topic revolves around Robot-Assisted Radical Prostatectomy, or RARP for short. This is a procedure where the surgeon removes the cancerous prostate out of the body using a robot. The prostate is a gland in the male reproductive system that is found under the bladder, with the urethra passing through it. It is responsible for producing seminal fluid that helps in transporting sperms [3]. Prostate is located beneath the bladder, the organ responsible for urine storage. The urethral tube is started from the bladder neck and goes through the prostate to the penis. It consists of a base, an apex, an anterior, a posterior and two lateral surfaces. Figure 2.1 illustrates the prostate anatomy in its two cases: benign and malignant.

Prostate gland cancer is the second leading cancer in men. According the US statistics for the year of 2020 published by American Cancer Society, there will be an estimate of 191,930 new diagnosed patients and 33,330 deaths [4].



(a) Benign prostate

(b) Malignant prostate

Figure 2.1: Illustrations of prostate anatomy in benign and malignant cases

Surgery is a one of the treatments for this disease along with other treatment plans such as active surveillance and external beam radiation. The type of surgeries that can be used to perform partial or full prostatectomy can be an open, a laparoscopic, or a robotic intervention. In most of the reputable hospitals in the world, including HMC, the state-of-the-art, FDA-approved surgical robot that is widely used is DaVinci Surgical

System by Intuitive Surgical [5]. The system comprises of three components. Figure 2.2 illustrates the different components.

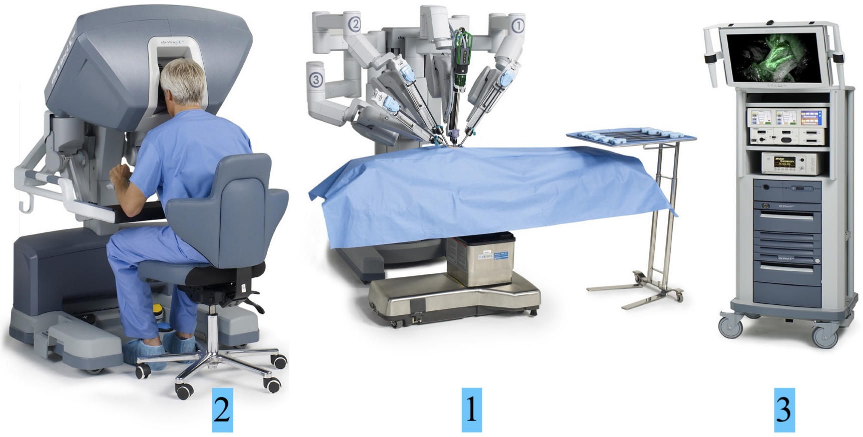


Figure 2.2: DaVinci Robotic system components

1. Patient Cart: a component where the patient lays down that is used to insert instruments through the patient's body.

2. Surgeon Console: a separate component where the surgeon sits and controls all instruments through intuitive hardware interfaces.

3. Vision Cart: a component that facilitates the communication across systems and displays live feed of the procedure for people attending the operation.

Robotic surgeries is a type of surgical interventions that is done using a robot. It is a minimally invasive procedure that has proven to have better surgical outcomes as oppose to other types of surgeries like open or laparoscopic interventions. RARP is an example of it. The procedure is done robotically, which means that the surgeon reaches the organ through small incisions made on the patient's abdomen. Surgical endoscope and instruments are then inserted in and the surgeon performs the entire procedure through them. By having few small incisions as oppose to an open surgery, patients experience less blood loss, fewer complications risks, less hospitalization and ultimately, better surgical outcomes. From a surgeon's perspective, robotic surgeries offer the surgeons the ability to be more meticulous while performing the surgery, especially when dealing with delicate areas that are surrounded by numerous nervous tissues like the male pelvis region. The robot offers better perceived ergonomics, safety, high-definition stereoscopic display, display magnification, tremor filtration, movement scaling, and wristed instrumentation with 6-degrees of freedom [6] [7].

Undergoing training for RARP is crucial for novice and intermediate surgeons to practice and assess their surgical skills before moving to the operating rooms. RARP is a procedure that involves many steps including urethra dissection step [8]. In this step, the surgeon dissects the urethra to free the prostate from its surrounding anatomy and extract it out of the body. The significance of the step lies in accurately determining the optimum cutting range or location on the urethra. Generally, prostate cancer lies in the prostate. However, there is always a chance of metastasis, meaning cancerous cells migrating to nearby healthy cells such as urethral tube. Incorrect dissection will compromise the oncological principals as the surgeon would leave cancerous tissues behind, which will consequently lead to positive margin. Positive margin is hypothesized to lead to an increased risk of biochemical recurrence of cancer, which defeats the purpose of the surgery of radical extraction of cancer. Additionally, poor dissection decisions, such as not leaving enough urethral stump for suturing after dissection, may affect patient's quality of life such as urinary continence and erectile dysfunction post-operatively [9].

2.2 Technical Background

This section will elaborate on the technical background needed for this thesis.

2.2.1 Convolutional Neural Network

Convolutional Neural Network, or CNN for short is a sub-type of Neural Networks that is widely used in the the field of computer vision locate a region of interest. Among the various applications, we find image segmentation such as in [10], [11], [12], image and video object detection such as in [13],[14],[15], [16], medical image analysis such as in [17], [18], [19], [20], etc. The name is inspired from the mathematical operation used throughout the network called "Convolution". In the context of images, a convolution is simply applying a filter (or kernal) to the input image to produce a feature map that is fed to the next layer in the network. Kernal is usually a 2D matrix that applied to the 2D input images and the result is a feature map. Convolutions are repeated throughout the network along with pooling layers. Pooling layers are responsible for downsampling the feature maps and reduce their spacial size to reduce the network computations and parameters, while maintaining the most important features intact. Max pooling and average pooling are the most used approaches. Simply put, rounds of convolutions and pooling break down images into features and analyze them.

The last major component of CNN is Fully Connected Layers (or FCL) is responsible to convert the output of convolution/pooling rounds into a tangible output in terms of probabilities, class predictions, regressions, etc. This is done by flattening the convolution/pooling output and turn them into a single vector and applies weights to output predicted classes and their probabilities. An illustration of a typical CNN architecture in Figure 2.3.

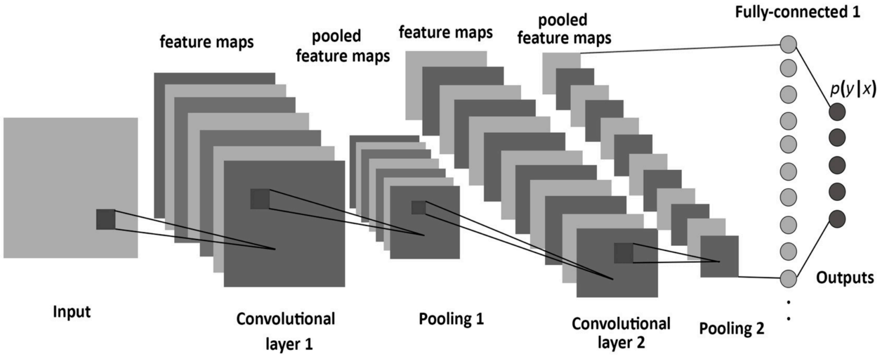


Figure 2.3: A typical CNN architecture

In this thesis, U-Net architecture is used to perform semantic segmentation tasks. U-Net is an encoder and decoder-architecture network that was intended specifically for medical image segmentation tasks. It consists of two major parts: contraction part (encoder), where the network digests the input image and learns deep features. The expansive part (decoder), where the network upsamples the learnt features and attempts to reconstruct a meaningful segmented image/mask with the same size of the input image. Based on the original paper that launched U-Net [21], authors report very good results using very small dataset sizes. Figure 2.4 shows the original architecture of U-Net with 23 convolutional layers. Blue boxes denote a series of 2D convolutional blocks followed by Rectified Linear Unit (ReLU) activation function. Red arrows denote a max pooling convolution with 2 X 2 sized-kernal to perfom down-sampling operations. Downsampling causes the feature channel to double its size by a factor of 2. To reverse the process, the expansion path upsamples the feature maps and a 2 X 2 convolutions that decreases the number of feature channels by 2. The feature map is concatenated with the corresponding cropped feature in the contracting part followed by a 2 X 2 convolutions and a ReLU. The output layer is 64 X 1 feature vector using a 1x1 convolution to map the different output classes.

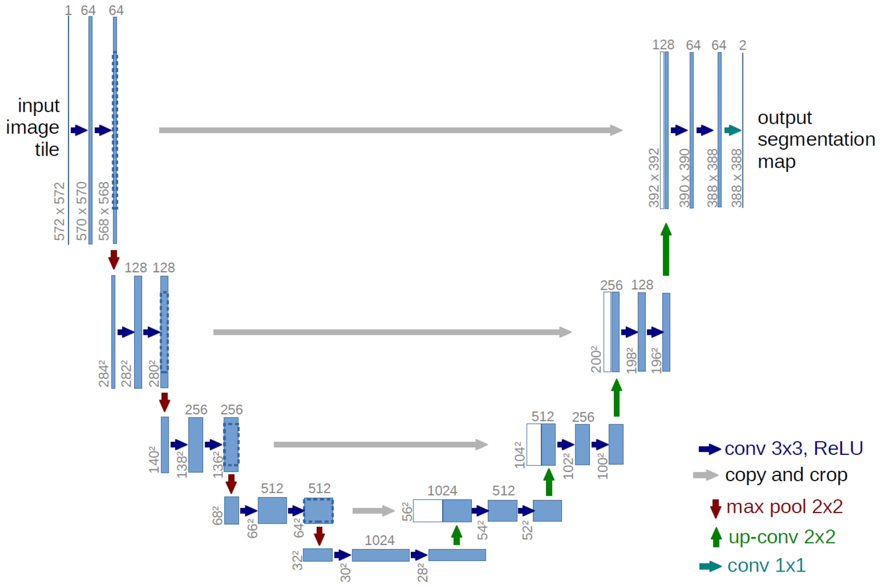


Figure 2.4: U-Net encoder-decoder architecture [21]

2.2.2 Medical Image Segmentation

Image segmentation is the process of partitioning the images into multiple parts, called "segments". These parts can be different depending on the application such as edge detection, thresholding, region-based segmentation, boundary detection, etc. For this task, deep learning is the approach of choice as it requires less human intervention during the training process. In traditional image processing mothods, a step of manual feature engineering is required before learning [22]. This would consequently add a burden of clinician intervention prior to the learning process, which is not suitable for this project. This is because clinicians are always busy and an automated feature extraction method is required. Additionally, computational resources for training purposes are not a constraint in this project. Post training, the model will be exported and used on a standard specifications computer to classify new images.

Image segmentation has been used in various applications including autonomous driving cars [23], [24] satellite image analysis [25], [26] medical image analysis [17], [18], [19], [20], etc.

When it comes to the medical field, image segmentation has gained an unprecedented attention to its potentials in solving medical imaging problems from segmenting anatomical structures such as in [27] [28] to segmenting surgical tools such as in [29] [30], kinematics-based and video-based surgical skills assessment such as in [31] [32] and automated frames extraction based on surgical step such as in [33] [34].

Novel image segmentation techniques has proven to be highly accurate in analyzing quite complex and fine medical images, including patterns that are not easily distinguishable with human eye. To show an example, Figure 2.5 (a) represents a slice from a brain volumetric image. The brain has several fluids the algorithm is attempting to segment. To perform the segmentation, each pixel in the original image will be assigned to a different label and create a mask accordingly. Thus, each pixel of the generated mask will have a label. This would enable the image processing unit to focus only on the region of interest (i.e. regions with specific labels) and ignore regions where no significant information is present.

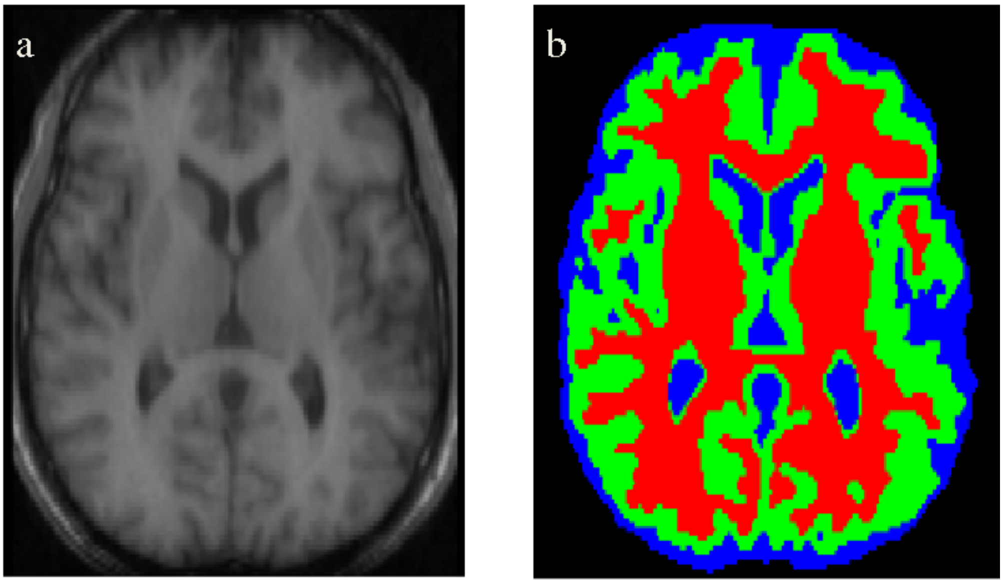


Figure 2.5: (a) A 3D volume slice of a brain image (b) Segmented brain parts using a statistical pattern recognition methodology. Red is the "white matter", green is the "gray matter" and blue is the "cerebrospinal fluid". [35]

CHAPTER 3: LITERATURE REVIEW AND FUNDAMENTALS

This chapter sheds the light on the similar existing works by fellow researchers and presents the key findings, similar to thesis work.

3.1 Medical Image Segmentation

Medical Image Segmentation has been widely used in last few years due to its powerful applications mainly in automating the detection process of anatomical structures, abnormalities, surgical tools, identifying surgical phases, etc. In similar works that attempt to segment the urethra using medical imaging modalities, in [36], authors attempted to segment out the male anterior urethra corpus spongiosum to assist urologists and surgeons treating urethral stricture disease. To do that, urologists annotated ground truth masks from ultrasound images, and using neural networks, they were able to reach a validation accuracy of over 90% in 2 different experiments. To evaluate the results, they have used similarity index, False-Positive Rate and False-Negative Rate as evaluation metrics. In a similar work, authors in [37] attempted to segment the urethra from a volumetric transperineal ultrasound of the female pelvic floor using 3D CNN, particularly HighRes3DNet. Ground truth data consisted of 35 ultrasound volumes. For model evaluation, Hausdorff distance and Dice coefficient were used. Standard Hausdorff Distance yielded in an average of $7.56mm \pm 1.65mm$ and a dice score of $0.65mm \pm 0.08mm$. Another similar application is lung segmentation [38]. In this work, authors attempt to segment out the lung then detect, localize and quantify lesion areas on COVID-19 patients. They used 51,027 computed tomography (CT) slices. Authors used several networks with different variants of DenseNet and ResNet as encoders. U-Net with DenseNet 161 encoder showed the best results for lung region segmentation with Jaccard Coefficient of 95.10% and Dice coefficient of 97.19%

Another application for medical image segmentation from surgical videos or images is extracting or highlighting anatomical ROIs for further processing such as colon polyps to detect colorectal cancer, which is the 3rd leading cancer cause of death after lung and breast cancer worldwide [39] [40]. For instance, authors in [39] attempted to detect, segment and localize polyps in a colonoscopy procedure. As dataset, authors used around 1000 images of high-resolution electromagnetic imaging system. For the segmentation task, authors have compared the performance of multiple CNN architectures such as U-Net, ResUNet, ResUNet++, HRNet, etc. Loss functions used were cross-entropy and dice loss. To evaluate the different models' performance, authors used Jaccard Coefficient, Dice coefficient, F2-score, precision, recall and overall accuracy. Results show that U-Net with ResNet34 backbone shows the best results among 9 others, with a Jaccard coefficient of 81%, Dice coefficient of 87.57%, precision of 94.35% and an overall accuracy of 96.81%. Another similar work for colon polyp segmentation [40] is where authors have used transfer-learning based segmentation with U-Net. They used CVC-ClinicDB, a dataset that consists of standard definition colonoscopy images. U-Net was the architecture of choice, and experimented different backbones. Accuracy, Jaccard coeffcient and Dice coefficient were used an evaluation metrics. U-Net with DenseNet169 backbone showed best results with 99.15%, 90.87%, 83.82% for accuracy, Dice and Jaccard, respectively. The next best scores were using U-Net with nceptionResNetV2 with 99.1%, 90.42%, 83.16% for accuracy, Dice and Jaccard, respectively. Another polyp segmentation work [41] used a deep encoder-decoder network. Authors used CVC-ClinicDB for training and validation and another dataset, ETIS-LaribPolypDB, for testing. For the evaluation, accuracy, Jaccard coefficient and Dice score were used.

Promising results were achieved with 0.975, 0.829 and 84.25%, respectively.

Although there exist several works for image segmentation using U-Net, Polyp segmentation works were the main inspiration to attempt trying image segmentation approach as they have proved to be effective with videos and images, which are the type of data we are using for the thesis.

Images segmentation for surgical videos is widely used in the literature, especially for surgical instruments segmentation and surgical phase recognition and identification. Surgical instrument segmentation gained increasing popularity due to its applications. Its applications vary from surgical workflow optimization, intraoperative reminders, objective skills assessment, etc. In [37], authors used RASNet, a U-shaped network, to segment and identify different surgical instruments. Jaccard index was used as a loss function. The model achieved 90.33% mean IoU score and 94.65% mean Dice score. Another similar work [42] that used U-Net, precisely U-NetPlus architecture to segment surgical instruments from laparoscopic images showed good results. Authors used Jaccard loss function and Dice as a performance metric. The model achieved 90.20% dice coefficient for binary segmentation, 76.26% dice in segmenting instruments parts and 46.07% dice in identifying instrument type. In a similar work, authors in [43] proposed a new architecture to segment surgical tools using RNN, and evaluated using Jaccard coefficient or IoU, specificity, sensitivity and balanced accuracy. Results show that the proposed method outperformed FCN-8 and ResNet-101 with 90.4% accuracy, 98.8% of specificity, 82.7% of IoU and 93.3% of balanced accuracy. Other works such as [29] and [30], authors use instance-based segmentation instead of pixel-wise segmentation to identify surgical tools using novel architectures. In [30], authors propose a new architecture called ISINet, in which it superseded the state-of-the-art performance for the benchmarked datasets EndoVis 2017 and EndoVis 2018. [29] proceeds to not only segment the surgical tool but also identify it. They are using Region-based Convolutional Neural Networks, trained with 333 images. For the evaluation metrics, authors used Intersection over Union (or Jaccard) and average precision. Results show 81% accuracy for at least 50% IoU.

3.1.1 Loss Functions

Working with machine learning or deep learning models necessitates the use of appropriate loss functions. Loss functions are functions that calculate the error between the actual labels and predicted labels. It is used as a metric to judge the goodness of a prediction in a real number. The loss values are meant to be minimized as possible. The loss value back-propagates through the network to learn from its wrong predictions and try to enhance in during the following forward-propagation. Simply put, the role of the loss function in machine/deep learning models is to optimize and decrease the loss value, to ensure maximum learning and minimum errors.

When it comes to the evaluation of image segmentation models, various loss functions are used in the literature to evaluate pixel-wise classifications including but not limited to binary-cross entropy, Jaccard Loss, Dice Loss, etc.

Binary-cross entropy is a loss function that is used in binary classification problem (in this case: C = 2; urethra and background). Equation 3.1 used often is below:

$$BCE = -\sum_{i=1}^{C=2} t_i log(f(s_i)) = -t_1 log(f(s_1)) - (1-t_1) log(1-f(s_1)) \tag{3.1}$$

In the context of this thesis, a segmented urethra usually represents from around 5% to

25% of the image. Similarly, white mask pixels would represent around 5% to 25%, and the rest of pixels are labeled as background. This cause class imbalance, which is a problem that is found vastly in the literature [44]. This means, the majority of the image would be 75% to 95% of the image would be labeled as "background", and the rest would be "urethra". While using Binary-cross entropy as a loss function, a natural behavior of this loss function is that the classification accuracy would be quite high (more than 90%). However, visually, predicted mask would be black and very far from being accurate. High value of accuracy comes from the fact that most of the pixels (i.e. background pixels) are correctly classified on the predicted image. However, Region of Interest (ROI) pixels are not. To overcome these limitations, other loss functions are usually used with imbalanced images segmentation problems such as Jaccard Loss and coefficient and Dice loss and coefficient.

Jaccard Loss, referred to also as Intersection Over Union (IoU) or Jaccard Similarity index is a metric to calculate the ratio of the intersection over union between two sets. As mentioned in [43], IoU (or Jaccard Loss) is one of the best segmentation metrics as it penalizes over and under segmentation. Jaccard loss function and metrics are used to calculate the percentage of overlapping pixels between the the ground truth mask and the predicted mask. Jaccard score varies from 0 to 1, with 0 meaning no overlapping and 1 meaning perfect overlapping. Equation 3.2 and Figure 3.1 illustrate Jaccard concept. In binary classification, positive and negative classes are to be distinguished. This leads to 4 different possible outcomes. Assuming that A is the positive class and B is the negative class.

- True Positive (TP): Object belongs to class A and classified as class A.

- False Positive (FP): Object belongs to class B and classified as class A
- True Negative (TN): Object belongs to class B and classified as class B
- False Negative (FN): Object belongs to class A and classified as class B

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| + |A \cap B|} \tag{3.2}$$

$$J(A,B) = \frac{TP}{TP + FP + FN} \tag{3.3}$$

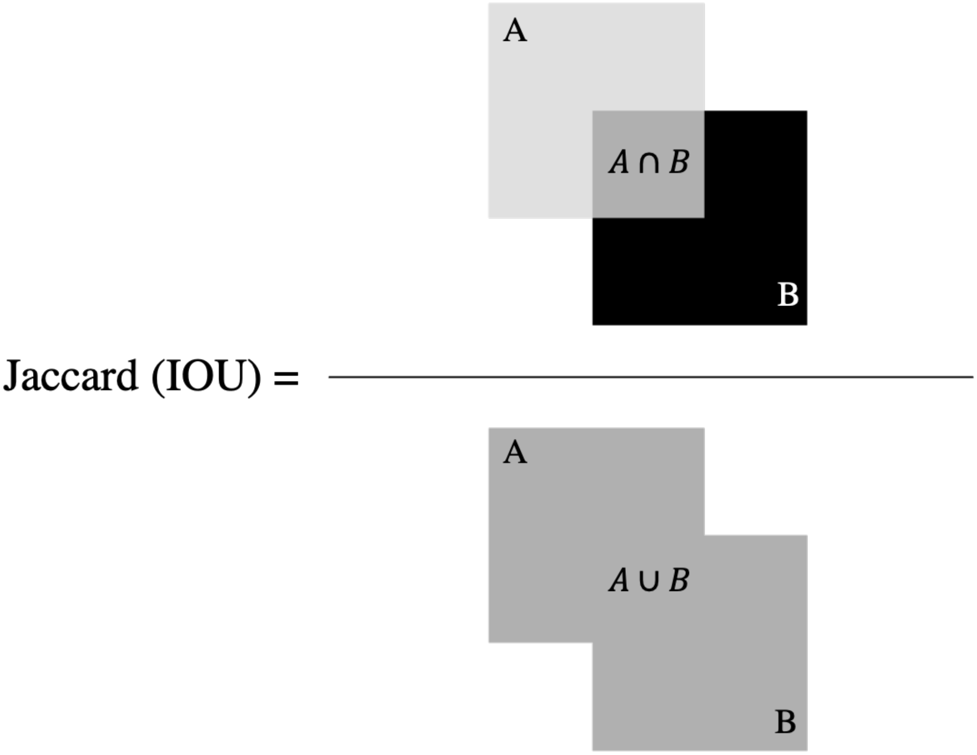


Figure 3.1: Illustration for Jaccard Coefficient or IoU

Sørensen–Dice coefficient, or Dice Coefficient for simplicity, is a metric that is used to calculate the similarity between two 2 samples or, images, in our case. The main difference from Jaccard index is that Jaccard only counts true positives once, unlike in Dice where it counts it twice in the numerator and denominator. Similarly, values of dice coefficient range from 0 to 1, where 0 indicates no overlapping and 1 indicates perfect overlapping. Equation 3.4 shows its formula.

$$DiceCoefficient = \frac{2TP}{2TP + FP + FN} \tag{3.4}$$

3.2 End-user Usability Evaluation

As robotic surgery is gaining more and more attention and being used heavily for intervention, there is an unprecedented need for training on these platforms. Recent studies that focus on robotic surgery training have shown that research is heading towards innovative and specialized training that enhances robotic surgical education and improves surgical skills. Integrating such tools helps to deepen robotic skills proficiency, according to [45]. It would require validating their effectiveness and use with experts and intended end-users to prove their usefulness and integrate them into training curricula. Such validations can vary depending on the nature of the prototype.

Questionnaires and surveys are among the most commonly used methodologies in the literature to capture users' feedback. There are numerous standardized surveys that can be readily employed and others that require adaptation to the application. Content validity is a commonly used evaluation method for validating interfaces and simulators and judging their appropriateness as a teaching modality. Such evaluation is important to be conducted to make sure it is delivering the concepts that it is meant to deliver. It is also important to perform incremental validations to make sure that a clinical prototype is being built rightfully and according to the clinical expectations. Content validity is heavily used in the literature. For this thesis, we focus the literature search for content validity on medical robotic and laparoscopic simulators. Robotic procedures are more and more in demand. Authors in [46] show that in a study between 2012 and 2018 among 169,404 patients across 73 hospitals, robotic surgeries surged for from 1.8% to 15.1%. Consequently, robotic simulators' demand is one the rise too. To validate their ability to teach, content validity is conducted, usually along with some other related tests such as face and construct validity. This is seen in [47], [48], [49], [50], [51] and [52] where usually the study is conducted with participants with different expertise. They are presented with pre and post study questionnaire and requested to judge the simulator using Likert scale subjectively. Some authors record other system objective metrics like time taken to finish a task, number of errors, etc. Once statistical analysis results are satisfactory, the prototype is considered validated. Laparoscopic surgeries quite similar to robotic ones. Validating simulators for such interventions is similar. In these studies [53],[54], [55], [56] and [57] authors validated laparoscopic simulator prototypes using the same approach for robotic simulators. In short, content validity is an effective and widely-adopted tool in the literature used in validating the usefulness of training simulators as a teaching modality and thus, will be used to evaluate this thesis's interface.

On the other hand, and since the user is intended to use the final prototype, it is critical to gather end-user feedback not only on its effectiveness, but also on its usability and satisfaction while using it. This would ensure that users feel involved in the development phase, excited to use the system, and prevent it from being unused once deployed. To quantify this, the Questionnaire for User Interaction Satisfaction (QUIS) is a questionnaire that is deemed to be potentially applied across different types of e-Health tools [58]. Although there exist many standardized usability questionnaires, QUIS seems to be the most comprehensive as it covers many aspects of interface usability. It is used to assess clinical applications types such as telemedicine applications. QUIS was used in [59], [60] and [61]. Although it is not not quite used in the literature for clinical applications (vastly used in non-clinical ones), it is the usability questionnaire of choice as it is comprehensive and reflects many aspects related to users satisfaction.

CHAPTER 4: URETHRA LOCALIZATION ON SURGICAL VIDEOS IMAGES: A DEEP LEARNING APPROACH

This chapter discusses the different training approaches experimented with box and contour segmentation using U-Net architecture to reach the appropriate results, along with the discussions and observations.

4.1 Solution Overview

The proposed solution is a desktop app where the trainee surgeon sits and train on several cases. These cases come from the images extracted from surgical videos from the operating room. As first step, the trainee surgeons will be asked to train on the first module called Tutorial mode, where they will have multiple variables to explore to deepen their understanding about the different cutting ranges' decisions. Once the surgeons master these skills, they move to the second module which is the Evaluation module, where the urethra on the image is automatically segmented and identified for the surgeons to start suggesting their cutting suggestions based on the patient case presented. The module will then evaluate their performance using a score. The score is then displayed and another image is shown to train. The automated segmentation is needed to help novice surgeon identify the urethra as well as automatically locate it on any image. The automation is done to leverage the videos acquired from the OR and extract key frames that include urethra. Surgeons can also export their scores over time to their respective supervisors through a module named Reporting. Parts of Evaluation module and Reporting module are out of this thesis's scope. Further details on this is mentioned on the next chapter: chapter 5.

Figure 4.1 shows the different modes (in green) and they are interacting with the different other modules.

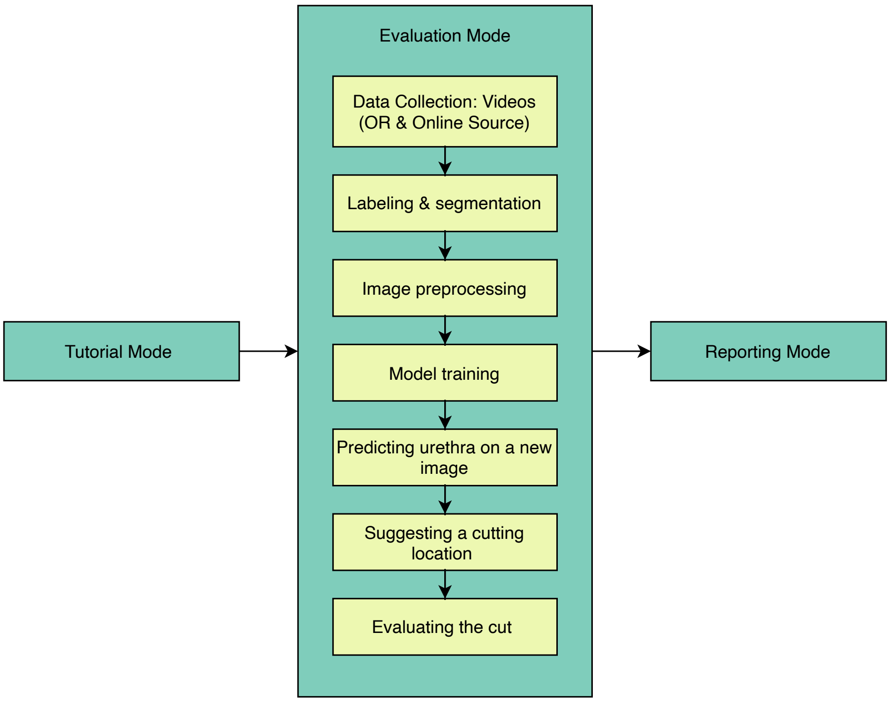


Figure 4.1: High-level thesis pipeline

Leveraging the potentials of deep learning with medical images, we attempted to apply it to the scenario of the detection of the ROI from 2D images extracted from surgical videos. The general automatic detection workflow is illustrated in Figure 4.2:

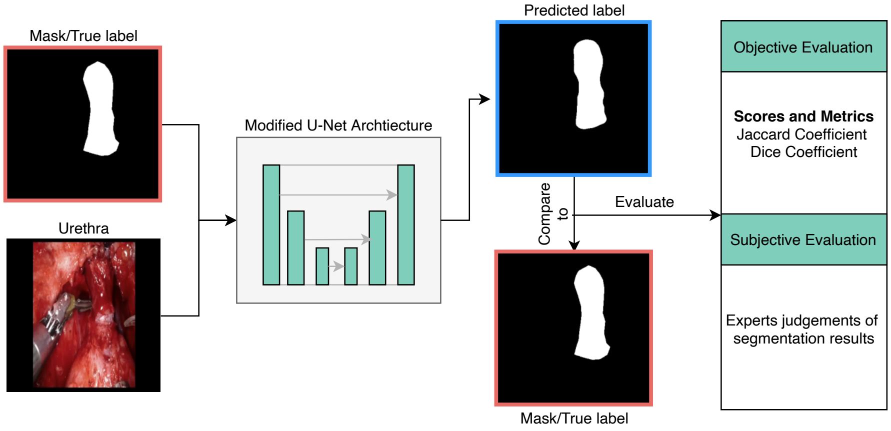


Figure 4.2: Workflow of the automated detection and evaluation of the urethra using U-Net

4.2 Data Collection & Labeling

The required data for this project are surgical images of urethra right before urethral dissection step. Such images are not readily available in a previous dataset. However, Hamad Medical Corporation's operating rooms (OR) have an archiving system in place that keeps videos of surgeries done robotically. Thus, four surgical videos were retrospectively collected from HMC's operating theaters and three from publicly available sources such as YouTube. OR videos were fully anonymized to keep patients' data privacy. With assistance of an expert urologist, videos underwent further refining process by determining video sections that have urethral dissection step. Then, frames were extracted from these videos using FFMPEG [62] library.

Out of three hours of surgical video footage, all collected videos have yielded in only 173 images/frames. Number of images with respect to length of footage seems to be below the expectations. This is due to several challenges in the footage such as:

1. Few RARP videos availability: Recording systems are not always set up and

thus, not all RARPs are recorded. This leads to fewer videos that can be used as datasets.

2. Some pre-urethral dissection step footage has bad quality due to different reasons. The main quality-degrading reasons are:

    - Smoke: During the procedure, smoke is commonly released due to tissue cauterization. This creates a foggy layer on the camera, and organs become hard to see. Figure 4.3 represents examples of surgical scenes collected.

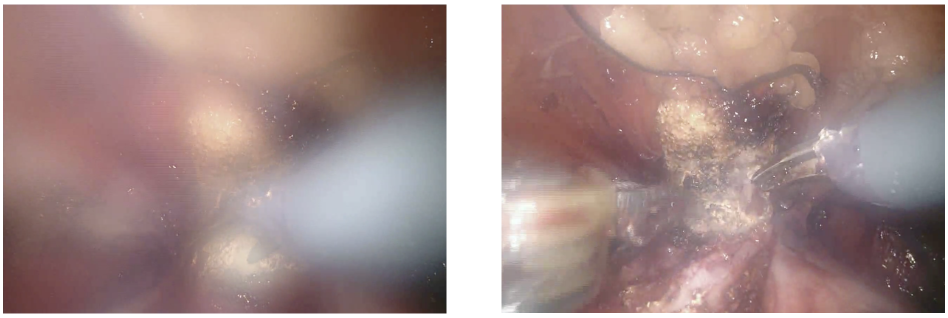


Figure 4.3: Surgical scenes with foggy view during tissue cauterization

    - Surgical tools: Surgical tools usually block the view as they are being used for cutting, holding, manipulating organs, etc. Thus, only few frames can be extracted where organs are not completely covered by them. Figure 4.4 shows examples of tools placed badly.

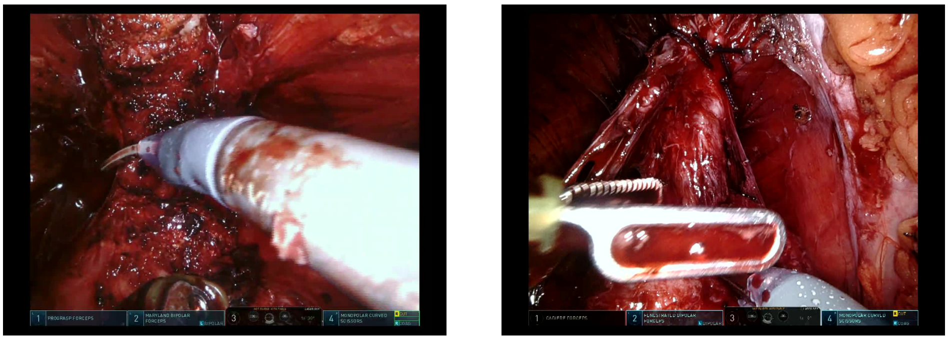


Figure 4.4: Surgical tools obscuring the vision from the urethra.

3. Urethrae images imbalance: Due to the various sources of images and the different lengths of the filmed urethra part in the videos, videos yield different number of images. Some of the videos generate very large images, while some others generate very few. This limitation would potentially result in the model learning being biased to the larger images collection.

4. High variability on urethra images: Urethra looks different in each video. This is because videos being recorded belong to different patients, at different ages and anatomies, different surgeons with different operating techniques, non-homogeneous illumination, different surgical endoscopes, etc. Below is Figure 4.5 that illustrates examples of how urethrae can differ from different videos:

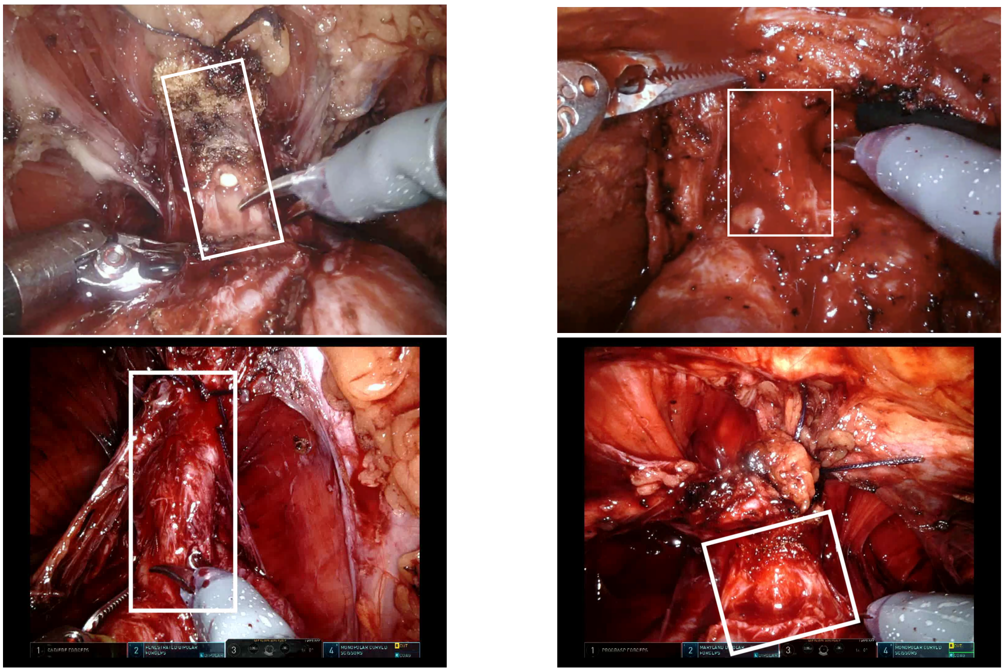


Figure 4.5: Urethrae (in white boxes) looking different from videos of different sources

To tackle the problem, image segmentation approach was used, where the original ROI on the image is further represented by a mask. The mask is usually a black and white image that has the same image size of the original one. The mask is all back, and the white spot (or spots) represent the pixels of the object or region of interest. Therefore, dataset was labeled by rectangular-shaped boxes that encloses the urethra. To label the data, RectLabel [63] is a tool that was used to localize the urethra on the images. This was done by drawing a box over the urethra in each of the images. Upon drawing the box, the annotator assigns it a class, which is "urethra" in this case. After annotation is done, all images are exported into Pascal VOC XML format, which where then converted to YOLOV3 annotation format:

```
<image path> <object-class> <x> <y> <width> <height>
```

A piece of code was then developed to digest the YOLOV3 annotations and covert them to boxes as figure 4.6 shows.

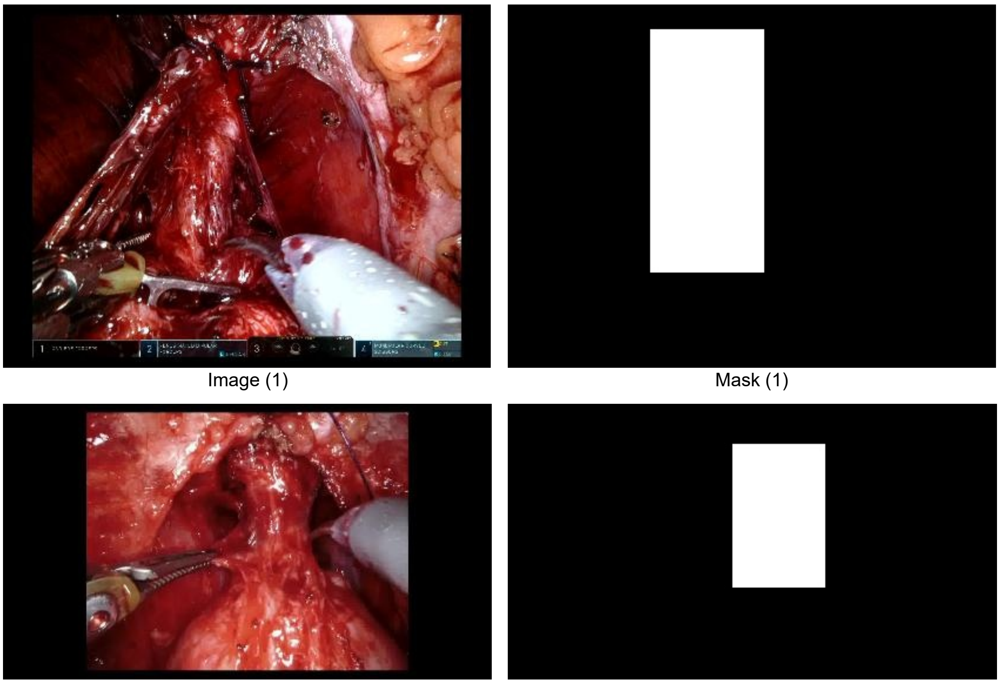


Image (1)

Mask (1)

Image (2)

Mask (2)

Figure 4.6: Example of two images segmented using box segmentation

Alternatively, urethra images were also labeled using another labeling techniques, which is segmenting the region of interest (ROI) exactly to its shape. In this case, we used splines to draw contours around exact edges of the urethra. To do that, we used an online, open-source tool named Computer Vision Annotation Tool or (CVAT) [64] to draw the splines and generate the masks. The figure 4.7 shows an example of the generated masks.

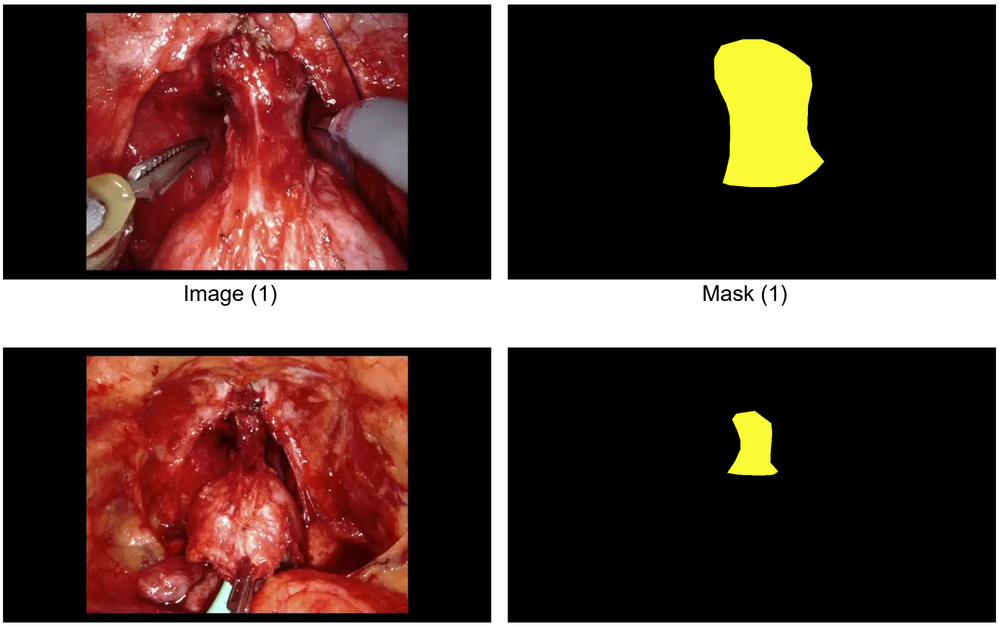


Image (1)

Mask (1)

Image (2)

Mask (2)

Figure 4.7: Example of two images segmented using contour segmentation

4.3 Data Pre-processing & Environmental Setup

Before training, data was augmented using the native Keras Image Data Preporcessing function `image_dataset_from_directory`. Images underwent pre-processing step by applying filters mentioned in Table 4.1. Examples of generated images are illustrated in Figure 4.8. In addition, training images and their respective masks underwent randomized shuffling before each epoch.

Table 4.1: Data Pre-porcessing filters applied on training dataset

Filter	Value
Rescale	$\frac{1}{255}$
featurewise_center	`True`
featurewise_std_normalization	`True`
rotation_range	90
width_shift_range	0.2
height_shift_range	0.2

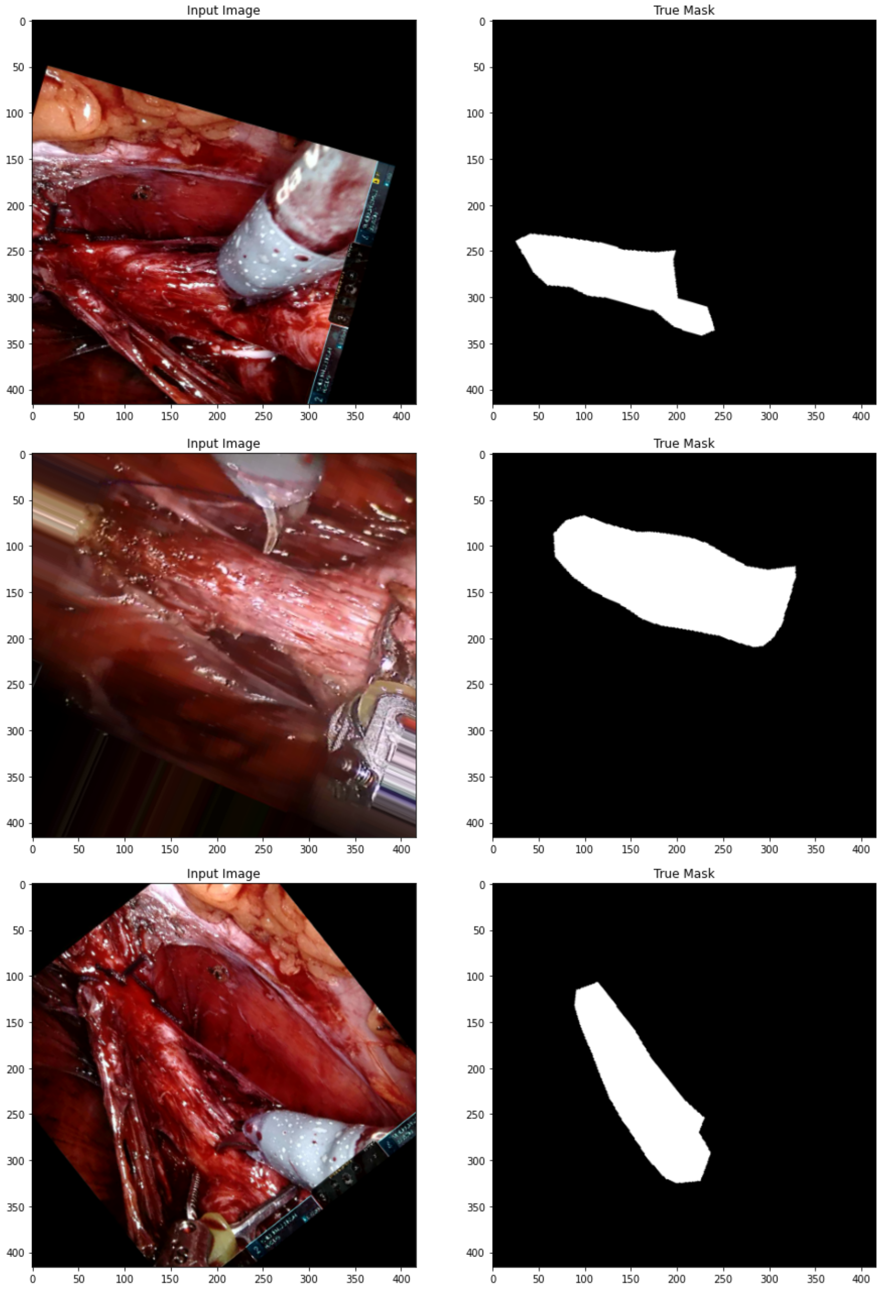


Figure 4.8: Examples of the generated data for training with filters applied

Total images used to train and test the model are 168 images: some images were deleted from the original 173 images due to clarity issues. These images were split into train-validation-test splits using 4 different variations. To ensure high randomization of data, each of the splits had 3 different versions, with a total of 12 versions of data. Figure

4.9 illustrates them. Each version has different images that is different from the other two versions. The main purpose of such randomization is to validate the performance of the model with different set of data. The intuition behind it is to prevent the model from overfitting over a certain set of data and ensure that the model behaves consistently over many variations of input data. This is because in machine learning in general and neural networks in particular, models attempt to learn quite complex relationships. The order of the data can be a feature that it learns, which is what we try to avoid to prevent overfitting. To add another randomization layer, `ImageDataGenerator` in Keras is used to generate batches of image data with real-time data augmentation. Argument `shuffle` was set to `True` for the training set. This enable to data generator to choose a new random rather than sequential set of images at each batch.

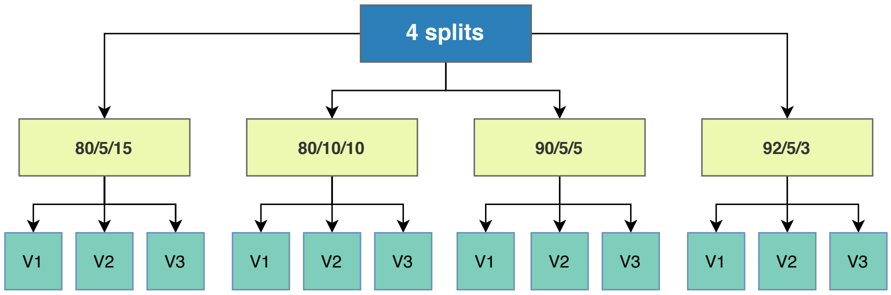


Figure 4.9: Different data splits used to train and validate the performance of the U-Net

Images were split as follows:

1. 80%/5%/15% (135-9-24): This split had 3 different versions different from each other. This will be referred to later as S1V1, S1V2, S1V3.

2. 80%/10%/10% (135-17-16): This split had 3 different versions different from each other. This will be referred to later as S2V1, S2V2, S2V3.

3. 90%/5%/5% (152-8-8): This split had 3 different versions different from each other. This will be referred to later as S3V1, S3V2, S3V3.

4. 92%/5%/3% (155-8-5): This split had 3 different versions different from each other. This will be referred to later as S4V1, S4V2, S4V3.

Training and testing sets were arranged in a specific format that is readily digested by U-Net. The below directory hierarchy shows the image splitting.

```
root
├─ train
│  ├─ img
│  │  └─ img
│  └─ mask
│     └─ img
└─ test
   ├─ img
   │  └─ img
   └─ mask
      └─ img
```

Training was done on Google Collaboratory platform, leveraging the free GPU capabilities for computationally-intensive tasks. Table 4.3 lists important environment requirements to run experiments. Different resources were assigned during different runs. The most common resources allocated are described in Table 4.2 [65].

Table 4.2: Google Collaboratory allocated resources of NVIDIA GPUs Tesla T4 & Tesla K80

Feature	Specifications
Graphics Processor	TU104
Memory size	16 GB
Bandwidth	320 GB/s
Cores	2560
Graphics Processor	GK210 x2
Memory size	24 GB
Bandwidth	480 GB/s
CUDA Cores	4992

Table 4.3: Google Collaboratory environment setup for experiments execution

Module/API	Version
Python	3.7.10
Keras	2.4.0
Tensoflow-gpu	2.1.0
opencv	3.4.2
numpy	1.18.1
matplotlib	3.1.3

4.4 Results & Discussions

For the segmentation tasks, U-Net was the convolutional neural network of choice as it was specifically designated for biomedical images from different imaging modalities such computerized tomography (CT) scans, Magnetic resonance imaging (MRI) scans and even Joint Photographic Group (JPG) images. U-Net showed good results with small dataset sizes and would be a good candidate to investigate its efficacy on surgical images, especially for an anatomy that is irregular in look, shape and sizes as demonstrated in Figure 4.5. Jaccard loss and Dice Coefficient are the the two most used metrics in the literature that evaluate the performance of image segmentation algorithms [27] [27] [39]

[38] [41]. They are usually chosen where there is a high imbalance in the in the masks between white and black pixels. Traditional metrics such as accuracy will give high accuracy score (more than 0.9) in the first few training and testing runs. This is an expected behavior as the model is indeed classifying most of the black pixels accurately and misses few of the white pixels.

4.4.1 Training with box-segmented images

For this approach, we tried a heuristic one that was inspired by YOLO3: a deep learning object localization network. The detected object is identified within a box. Several experiments were conducted to reach an acceptable training and validation results. The model was trained with 50 epochs, 100 epochs, 150 epochs and 300 epochs, respectively, while changing few hyperparameters such as the learning rates, optimizers, train and test batch size, U-Net levels, loss functions, etc. Loss function as such binary-cross entropy and mean-square error were used along with accuracy as a metric and they showed very high accuracy at the early stages of training, While testing, masks appear to be black. Figure 4.10 shows an example of the predicted masks while resulting in 98.42% accuracy. This is due to class imbalance. To mitigate this, Jaccard loss was used. However, results remained inaccurate. Figure 4.11 shows few examples of the predicted masks. At this point, an alternative solution was applied using contour-based segmentation.

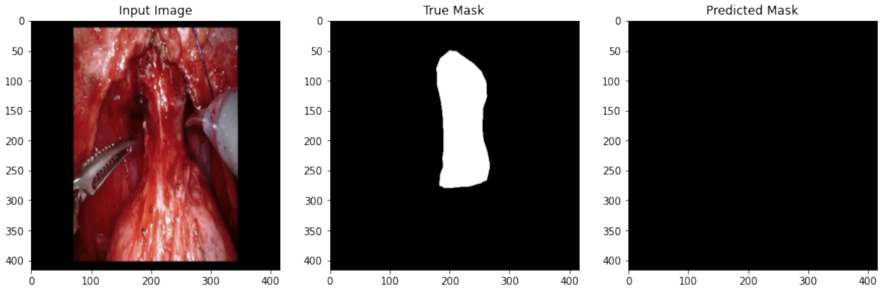


Figure 4.10: Predicted segmentation results showing high accuracy in metrics but no predicted masks.

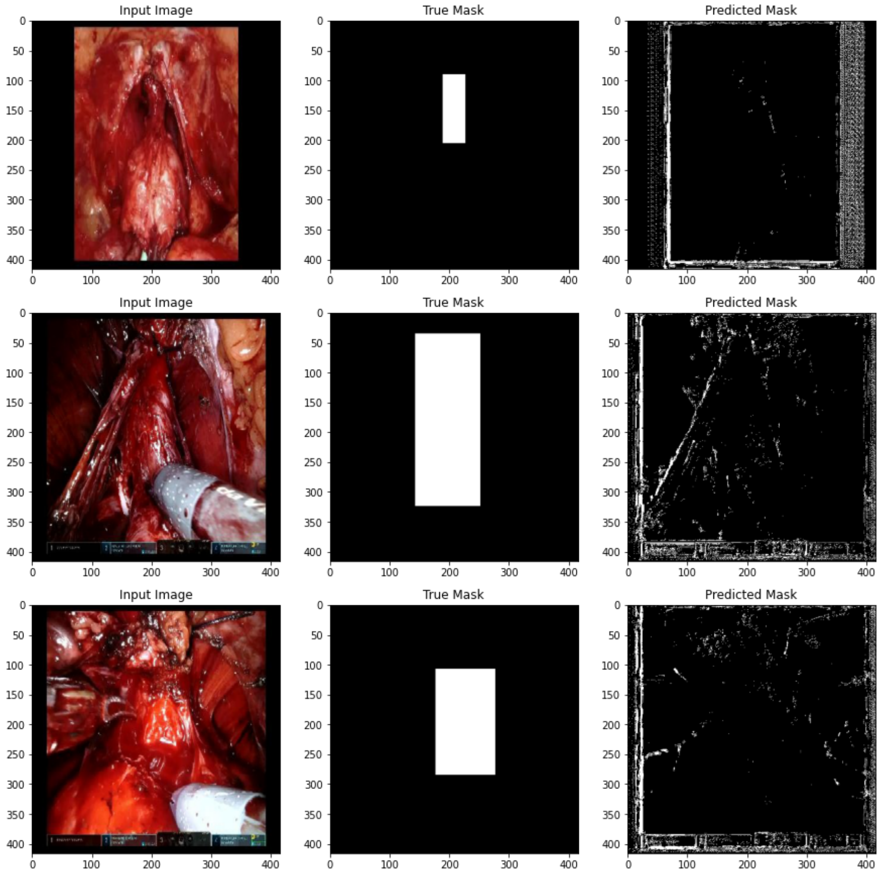


Figure 4.11: Predicted segmentation results using box-segmented images, using binary-cross entropy loss 100 epochs. As seen, predicted masks look very different from ground truth.

4.4.2 Training with contour-segmented images

In contrast with box segmentation, coutour-segmented images showed promising results. Total images used were 168, these images were split as illustrated in Figure 4.9 to determine the one one that yields best results and test the model's performance against them.

Image height and width were fixed: $416X416$ pixels. Adam was the optimizer used with $1 * 10^{-3}$, which showed no results. Thus, learning rate was lowered to $1 * 10^{-5}$. The loss function used was Jaccard loss function and the evaluation metrics used where Jaccard Coefficient and Dice Coefficient. Table 4.4 shows the different hyperparameters and report the one with the best results.

Figures 4.12 and 4.13 show the predictions along with true labels along with the training and validation trends for Jaccard coefficient, Dice coefficient and Jaccard loss.

As shown in figure 4.12, in model loss subplot, it is clear that training and testing loss values are behaving similarly. In fact, training loss is a bit higher than testing loss. This behaviour is expected due to the low number of testing set of images. Similarly, in model dice_coef subplot, training and testing coefficients are nearly perfectly converging, which indicates a predicted values are very similar to ground truth values. Another metric to assess the model model is jaccard_coef subplot. As shown, plotted values show similarity in the performance of unseen data, which is a good indication of the model's robustness on locating the urethra on new set of images.

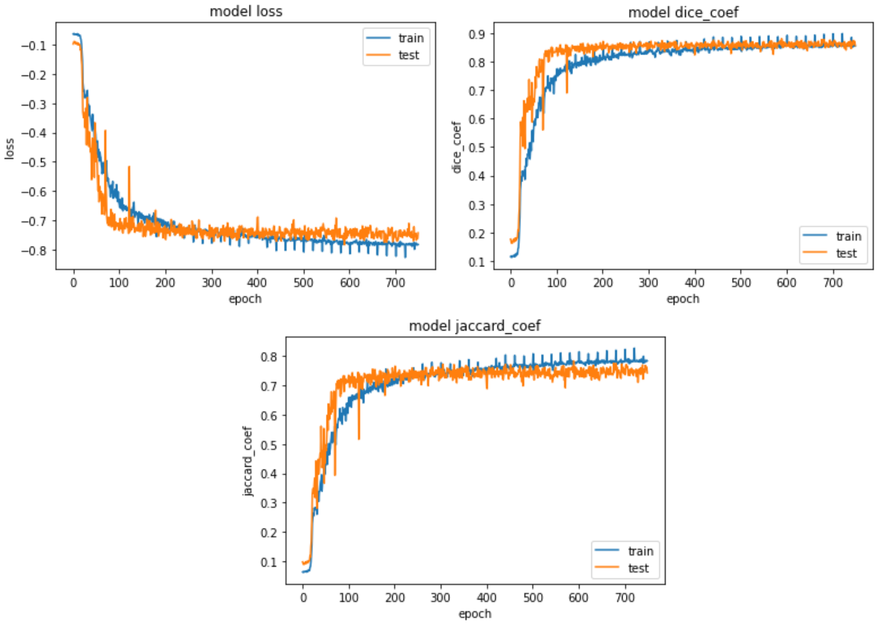


Figure 4.12: Trends of training and validation score of Jaccard coefficient, Dice coefficient and Jaccard Loss over 750 epochs

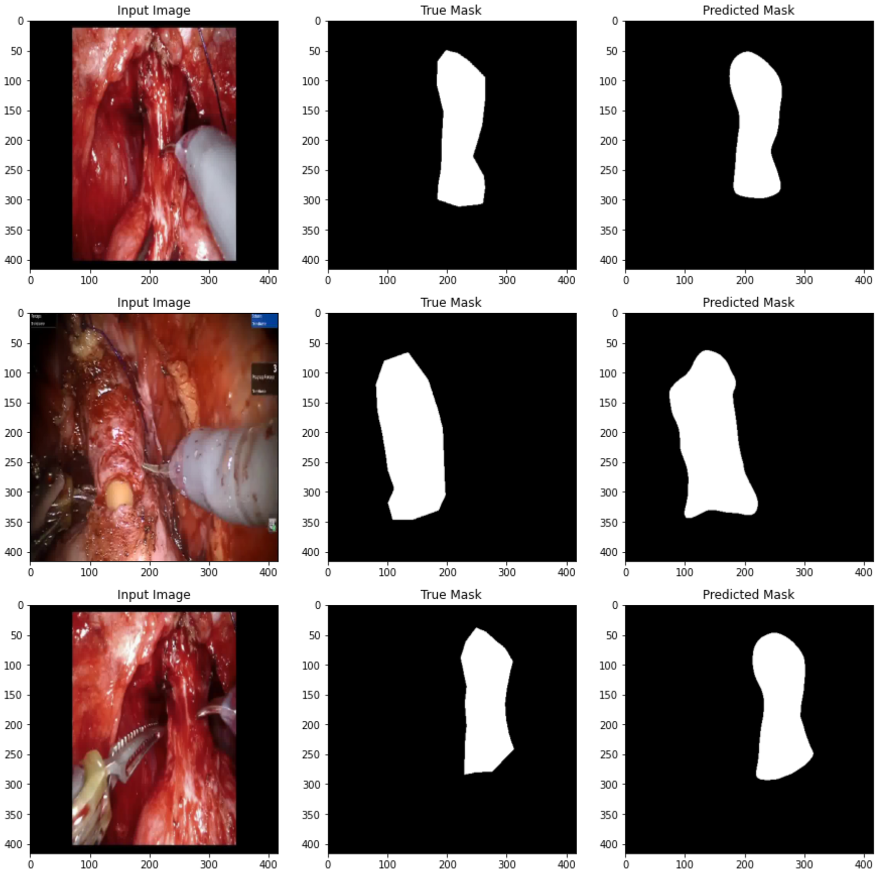


Figure 4.13: Results of the model with the best Jaccard Coefficient and Dice Coefficient scores: 78.42% & 88.11%, respectively

Table 4.4: Key parameters and results of the trained U-Net model, validated over test set

No.	Train-BatchSize	Test-BatchSize	# epochs	# of U-Net Levels	Best Val Jaccard Score	Best Val Dice Score
S1V1	8	8	550	6	0.74272	0.8471
S1V2	8	8	550	6	0.73659	0.8582
S1V3	8	8	550	6	0.2728	0.4245
S2V1	8	8	550	6	0.50555	0.6600
S2V2	8	8	550	6	0.76707	0.8682
S2V3	8	8	550	6	0.72524	0.8325
S3V1	8	8	550	6	**0.79161**	**0.8788**
S3V2	8	8	550	6	0.77387	0.8634
S3V3	8	8	550	6	0.7594	0.8396
S4V1	16	8	100	4	0.2728	0.4245
S4V1	8	8	250	6	0.7013	0.8288
S4V1	8	8	550	6	0.7619	0.8695
S4V1	8	8	750	6	**0.7842**	**0.8820**
S4V2	8	8	750	6	**0.7799**	**0.8807**
S4V3	8	8	750	6	**0.7805**	**0.8811**

Highlighted values in Table 4.4 represent the best results among all the experiments done. S3V1 showed highest Jaccard score while S4V1 showed highest Dice Coefficient score. To investigate this further, we checked the similarity of the training and testing data that could potentially yield to this high score. As expected, 80 out of 152 images were from the same video. All the testing images were also from the same video. This explains the higher score. Checking S4V1, training and testing images were highly randomized. Thus, we deduce that S4 in the split that generates the best results. This is because the model is training on more examples and able to generalize better.

To test the model's resilience, 72 additional experiments (6 identical experiments for each of the 12 splits) were conducted to show the inter-variability of the best Jaccard and Dice scores obtained across different trials. Table 4.5 shows the summary of the experiments by averaging the the best scores and deriving the standard deviation (STD). Looking at the results, the model showed high resilience, with a maximum STD of ± 0.2198. All experiments ran for 550 epochs, except for S4 splits.

Table 4.5: Mean and standard deviation of each split's best Jaccard and Dice coefficient values, ran for 6 times

No	Mean Best Jaccard		Mean Best Dice	
	Mean	STD	Mean	STD
S1V1	0.74312	±0.114	0.8499	±0.1930
S1V2	0.73892	±0.129	0.8533	±0.1229
S1V3	0.27451	±0.192	0.4253	±0.1875
S2V1	0.51219	±0.129	0.6681	±0.1911
S2V2	0.77139	±0.1823	0.8701	±0.1982
S2V3	0.73114	±0.1348	0.8291	±0.1368
S3V1	0.79108	±0.1391	0.8791	±0.1713
S3V2	0.79211	±0.2198	0.8622	±0.1422
S3V3	0.7605	±0.1332	0.8406	±0.1181
S4V1	0.7833	±0.0118	0.8891	±0.2073
S4V2	0.7783	±0.1722	0.8799	±0.1298
S4V3	0.7703	±0.1685	0.8871	±0.1762

After the generation of masks from the deep learning model, images were post-processed by applying the predicted mask over the original image to be able to integrate it in the Evaluation mode. The predicted urethra is surrounded with a box, allowing the surgeon to suggest locations. Figure 4.14 is an example: white mask is the true mask, yellow mask is the predicted mask, yellow box is drawn from the first white pixel found to the last white pixel found, and imposed over the original image.

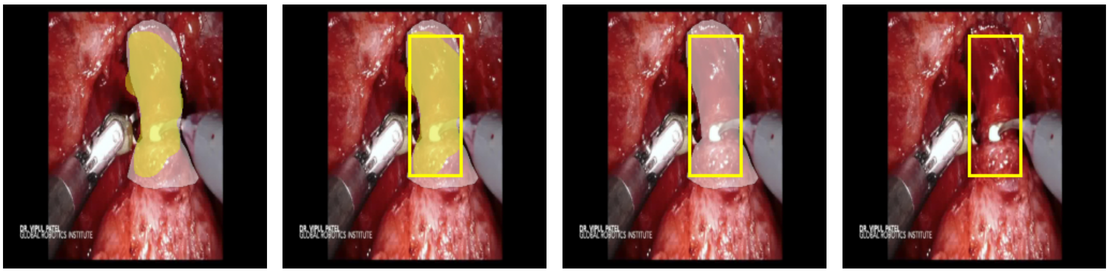

DR. VIPUL PATEL
GLOBAL ROBOTICS INSTITUTE
DR. VIPUL PATEL
GLOBAL ROBOTICS INSTITUTE
DR. VIPUL PATEL
GLOBAL ROBOTICS INSTITUTE
DR. VIPUL PATEL
GLOBAL ROBOTICS INSTITUTE

Figure 4.14: True and predicted mask and box drawn over original urethra image

Preliminary discussions with one expert urologist surgeon shows that the results are acceptable and this can be used for surgeons to learn to identify the urethra and suggest cutting locations on the urethra. Nevertheless, this needs to be further verified with more surgeons and thoroughly validated at a later stage. Due to the COVID-19 situation and the unavailability of surgeons, conducting validation studies is not possible.

CHAPTER 5: CLINICAL VALIDATION

This chapter discusses the efforts into validating the clinical needs for a training tool during for urethral dissection in a RARP procedure. Validations have been conducted before the development initiation and after it. It discusses thoroughly the methodology and key findings. A major part of this chapter has been accepted to be published in a manuscript entitled: "Content Validity and User Satisfaction Evaluation of Visualization Training Tool for Surgeons for Urethral Dissection during Robot-Assisted Radical Prostatectomy" in 2021 5th International Conference on Medical and Health Informatics (ICMHI 2021) in Kyoto, Japan.

5.1 Preliminary Clinical Need Validation

In this context, clinical need validation refers to the process of verifying that the problem in question is significant or relevant to surgeons to find solutions for it. This is done to make sure that the output of the project will have a significant impact on the field of surgical training. To do so, we have followed the Stanford University BioDesign guidelines for value-based innovation of medical applications. Based on the book's recommendations, a series of shadowing sessions and questionnaires were conducted to validate the clinical unmet need.

Training shadowing sessions are sessions where we accompany the robotic surgery trainer during his training sessions and we observe his training and skills assessment methodology. In HMC, trainees are required to pass a free online introductory course on DaVinci Surgery Community [66] to teach them the fundamentals of robotic surgery. After that, a trainee is moved to training on DaVinci Surgical robot that is dedicated for this purpose in HMC-Itqan, a state-of-the-art clinical simulation and innovation center.

After that, the trainee is requested to attend few robotic surgeries and assist in some of them, and gradually start performing simple surgical tasks.

To validate the clinical needs, three expert RARP surgeons were interviewed. The questions asked and their answers are available in Appendix A.

Based on their feedback, it is certain that this tool is quite useful both as an intra-operative and training tool. This validates the assumptions of this tool's usefulness in clinical practice. Their feedback will help steer the development into the right direction. Figures 5.1 and 5.2 summarize their feedbacks.

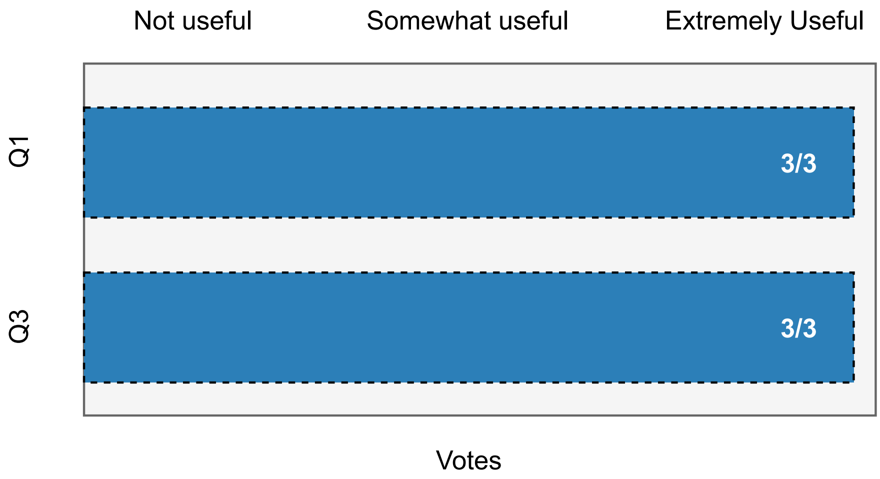


Figure 5.1: Surgeons feedback summary on Q1 and Q3

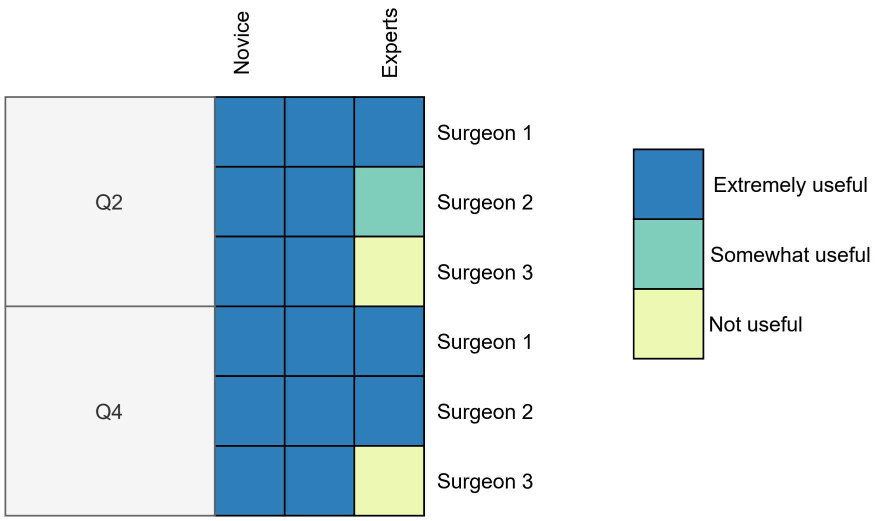


Figure 5.2: Surgeons feedback summary on Q2 and Q4

As part of early clinical validation, surgeons were interviewed about the usefulness of such training module. Some surgeons believe that, roughly, surgeons tend to always perform dissection in the middle of the urethral tube. Consequently, this envisioned training tool will not as useful. While researching, and based on two variables only, we found that middle dissections are true for only $\frac{4}{9}$ (44%) of cases. This is not true for the remaining $\frac{5}{9}$ (56%). Figure 5.3 shows the cutting decisions based on two variables: prostate size and cancer location.

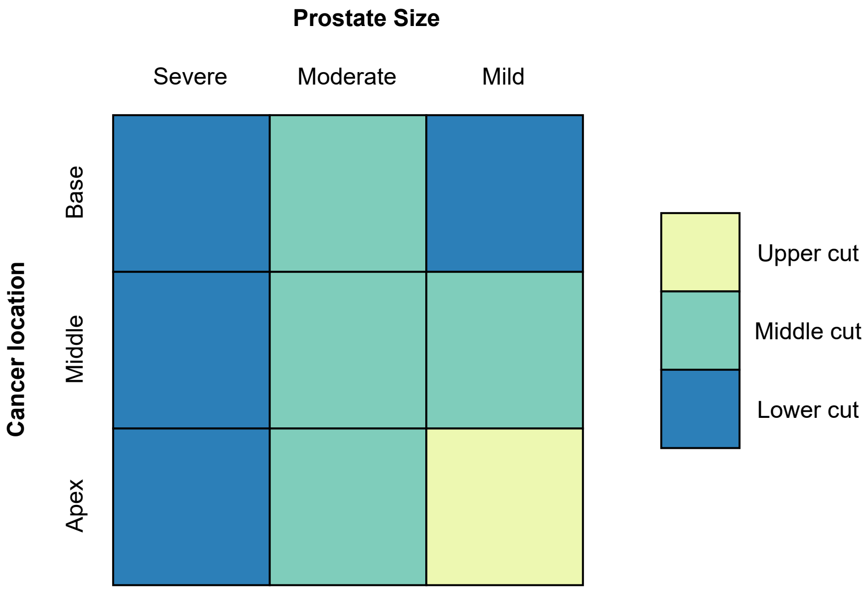
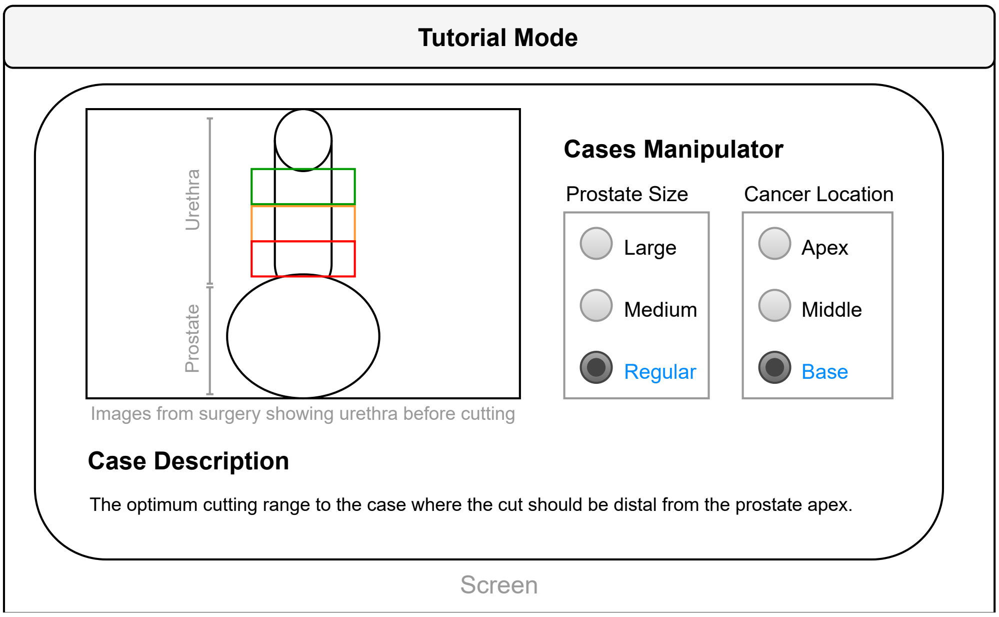


Figure 5.3: Matrix visualizing the different cutting decisions based on two variables: prostate size and cancer location

5.2 Interface Design

For easy integration between Python-implemented deep learning models outputs and the interface implementation, Python Tkinter version 3.8 module was used.

Before the implementation begins, several rounds of wire frames designs were co-designed and reviewed with an expert surgeon to ensure that information and their placement on the window are logical and conveys the information clearly and accurately. Figures 5.4, 5.5 and 5.6 show wire frames suggested for different modes.



Cases Manipulator

Prostate Size

○ Large
○ Medium
● Regular

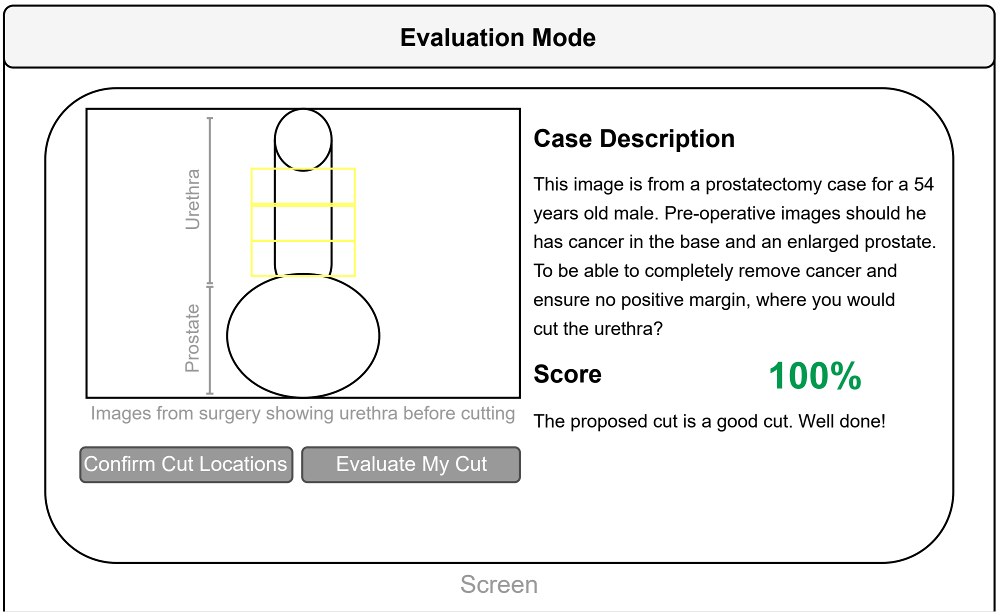
Cancer Location

○ Apex
○ Middle
● Base

Case Description

The optimum cutting range to the case where the cut should be distal from the prostate apex.

Figure 5.4: Wire frame design for tutorial mode



Case Description

This image is from a prostatectomy case for a 54 years old male. Pre-operative images should he has cancer in the base and an enlarged prostate. To be able to completely remove cancer and ensure no positive margin, where you would cut the urethra?

Score 100%

The proposed cut is a good cut. Well done!

Figure 5.5: Wire frame design for evaluation mode

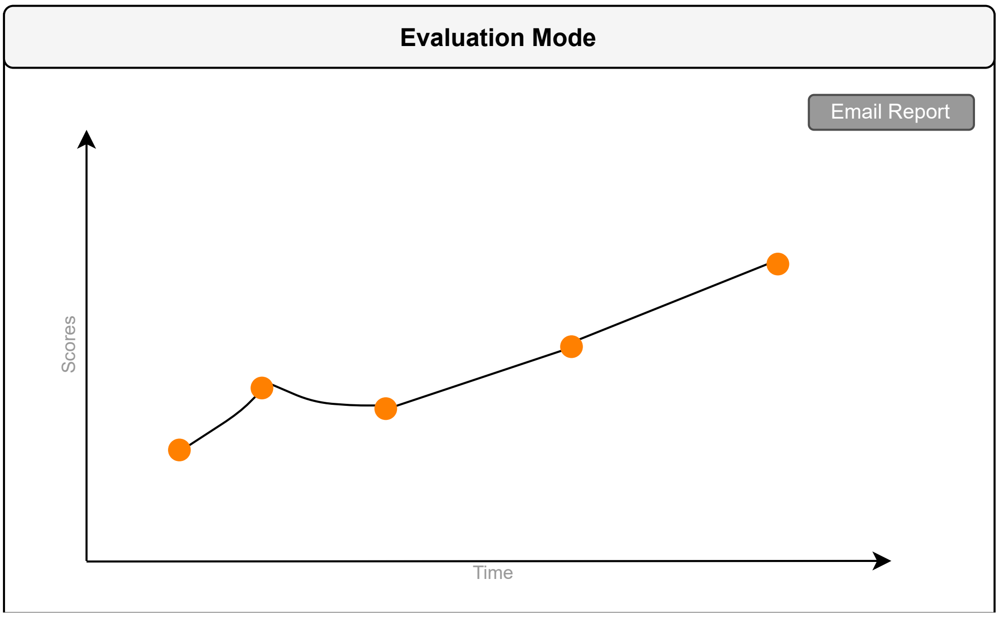


Figure 5.6: Wire frame design for reporting mode

The interface was designed into three modules. Each module serves a separate purpose: tutorial mode, evaluation mode and reporting module. These modules are further explained next.

5.2.1 Tutorial Mode

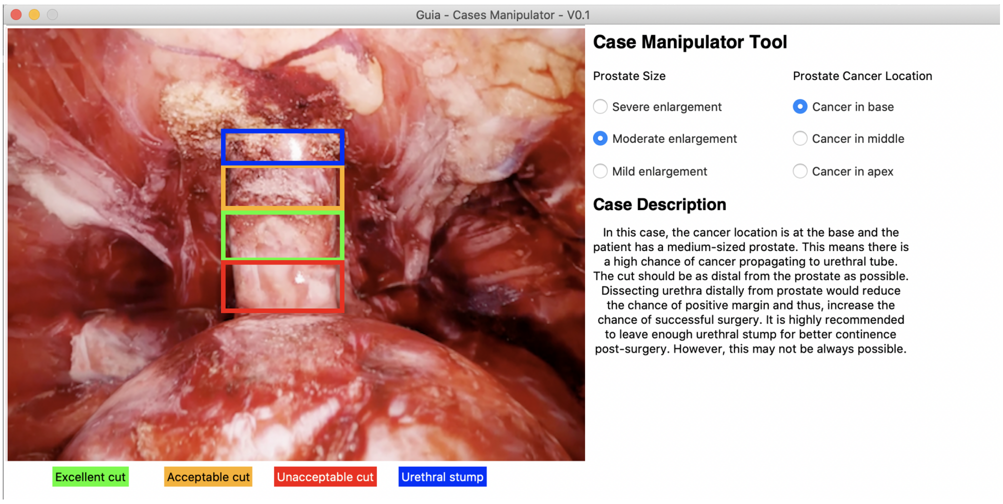
On the application, the surgeon will be trained on the optimum cutting range decision-making using the tutorial mode named "Case manipulator." Interface elements are:

Part 1: Manipulation options: the trainee is presented with nine variables: cancer location in the prostate (apex, middle, base) and prostate size (mild enlargement, moderate enlargement, severe enlargement). During training, the trainee chooses a combination: cancer location in apex and prostate size is mild, for instance.

Part 2: Optimum dissection regions: Once a combination is selected, colored boxes are displayed on the urethra to delineate the optimum cutting ranges along with case description text to explain further the dissection decisions. In the case of Figure 5.7, the optimum cutting ranges can be proximal to the top of the prostate. Additionally, boxes that are overlayed on the urethra are color-coded to mark different areas of accepted or unaccepted cuts. A green box contour denotes an excellent cut, an orange box contour denotes an acceptable but not preferred cut, a red box contour denotes an unacceptable cut, and a blue box contour denotes the urethral stump that needs to be taken into consideration while cutting, as emphasized in [67]

Part 3: Description: Once a combination is selected, a case description appears to elaborate further on the specific case.

Figure 5.7 shows a screenshot of the interface with the case of moderate prostate enlargement and base cancer.



Case Manipulator Tool

Prostate Size
○ Severe enlargement
● Moderate enlargement
○ Mild enlargement

Prostate Cancer Location
● Cancer in base
○ Cancer in middle
○ Cancer in apex

Case Description

In this case, the cancer location is at the base and the patient has a medium-sized prostate. This means there is a high chance of cancer propagating to urethral tube. The cut should be as distal from the prostate as possible. Dissecting urethra distally from prostate would reduce the chance of positive margin and thus, increase the chance of successful surgery. It is highly recommended to leave enough urethral stump for better continence post-surgery. However, this may not be always possible.

Figure 5.7: System graphical user interface (GUI) with a case example: moderate prostate enlargement and base cancer

5.2.2 Evaluation Mode

After mastering the tutorial mode, the surgeon is ready to evaluate their skills. In this mode, the surgeon will be presented with different urethra images extracted from intraoperative surgical videos. Surgeons can also input their own images to train on them. The machine learning module trained with urethra images will localize the urethra on an image in a box then the surgeon needs to suggest optimal cutting ranges based on the patient's case presented. The case is described in terms of prostate size and cancer location in the prostate. Once the surgeon submits the answer, the tool will automatically evaluate their performance and display a score accordingly. It will also show the correct answer and recommendations on how to correct their decision to improve the score on the next round. The score gets stored, and a new urethra image is displayed with a different case to train on.

For the automatic evaluation, and after the prediction of the urethra location using the deep learning model, the module would pick a random combination of prostate size and cancer location and present it to the surgeon as a case for him/her to decide where to cut. After the surgeon suggests a cutting location, the module will evaluate it.

5.2.3 Reporting Module

Assuming that each surgeon is being monitored and supervised by an expert surgeon, it is important to keep them informed of the trainee surgeon's performance. Thus, the system can record all scores for trainee surgeons, plot them over time, and send timely reports to supervisors in the form of email reports.

5.3 Tutorial Module Validation

We aim to validate the system using two validation methodologies to assess two different aspects of the prototype. Five urology surgeons were voluntarily recruited for the study with varying years of expertise from multiple Hamad Medical Corporation hospitals across Qatar. All surgeons have consented to the participation and the use of their feedback for further analysis. Table 5.1 shows the different surgeons' demographics. Please note that "0.5" means that a surgeon has expertise in two types of surgeries.

Table 5.1: Demographic profile of participants

Demographic Variable	Resident	Fellow	Consultant
Number of participants (n=5)	$1(20\%)$	$2(40\%)$	$2(40\%)$
Type of surgeries			
Robotic	$1(20\%)$	$2(40\%)$	$0.5(10\%)$
Laparoscopic	$1(20\%)$	$2(40\%)$	-
Open	-	-	$1(20\%)$
Endoscopic	-	-	$0.5(10\%)$
# of experience yrs	3	$6.5(6,7)$	$10.5(6,15)$
# of performed surgeries	300	$600(200,1000)$	$1250(500,2500)$

Recruitment of surgeons was done according to specific inclusion and exclusion criteria. A surgeon that is included in the study has to be a urologist with experience in robotic, laparoscopic, or endoscopic procedures. A surgeon who does not satisfy these requirements will not be qualified for the study. During the session, surgeons were briefed with an introduction about the unmet clinical need and presented with the prototype and its main features and were asked to interact freely with it. After the session, surgeons were given a survey and requested to answer all questions individually, with no discussion with other surgeons to ensure that they were unbiased. After collecting their feedback, the answers were statistically processed and analyzed. We consider a statement with a mean value of 7 to be weak and needing attention in the next iteration.

5.3.1 Content validity

Content validity is widely used in the literature to judge a prototype's appropriateness as a teaching modality. This is performed through a series of 8 questions that should be answered by experts. To answer the questions, participants were asked to rate a statement from 1 to 10. Table 5.2 lists the questions and their respective average scores.

Table 5.2: Content validity mean scores ratings by surgeons, out of 10

Content validity statements	Mean	Range	Experts mean	Non-experts mean
1. Useful for training residents	8	$6-9$	7.5	8.3
2. Useful for training fellows	7.6	$5-9$	7	8
3. Useful for expert surgeons	5.4	$3-6$	5	5.6
4. All novices to robotic prostatectomy surgery should undergo training on the module to learn where to cut the urethra prior to performing robotic surgery on patients	7.2	$5-10$	7.5	7
5. Case manipulator is well presented and help training surgeons explore different case combinations	7.6	$6-10$	8	7.3
6. Color-coding helps the trainee visualize the proper cutting ranges according to the patient's case	8.4	$6-10$	8	8.6
7. Cases explanation is important to describe further the cutting choices	8.4	$7-10$	8.5	8.66
8. Cutting ranges are well-illustrated using boxes in the urethra	8.4	$8-10$	9	8

5.3.2 Questionnaire for User Interaction Satisfaction

Questionnaire of User Interaction Satisfaction or (QUIS) is a ready-made questionnaire widely used in the literature and further adapted to the project's specifications. QUIS is used to subjectively assess the user's satisfaction with the interface. The survey contains a series of 27 questions. Participants were asked to rate a statement from 1 to 10 when filling out the questionnaire. Table 5.3 lists the questions and their respective average scores.

Table 5.3: QUIS mean score ratings by surgeons, out of 10

User satisfaction questionnaire statements	Mean
1. Overall reactions to the software	8.5
1.1 Terrible/wonderful	8.6
1.2 Difficult/easy	9.4
1.3 Frustrating/satisfying	8.4
1.4 Inadequate power/adequate power	8.4
1.5 Dull/stimulating	7.6
1.6 Rigid/flexible	8.8
2. Screen	9.3
2.1 Characters on screen (hard to read, easy to read)	9.8
2.2 Highlights on-screen simplify tasks (not at all, very much)	9
2.3 Organization of information on the screen (confusing, very clear)	9.2
3. Terminology and system information	8.5
3.1 Use of terms throughout the system (inconsistent, consistent)	7.8
3.2 Terminologies are related to the task (not at all, very much)	8
3.3 Position of messages on screen (inconsistent, consistent)	9.2
3.4 Computer keeps you informed of what you are doing (never, always)	9
4. Learning	9
4.1 Learning to operate the system (difficult/easy)	9
4.2 Exploring new features by trial and error (difficult/easy)	8.75
4.3 Remembering names and use of commands (difficult/easy)	9
4.4 Tasks are performed in a straight-forward manner (never/always)	9.2
5. System capabilities	8.86
5.1 System speed (too slow/fast enough)	8.4
5.2 System reliability (unreliable/reliable)	8.8
5.3 System tends to be (noisy/quiet)	9.6
5.4 correcting your mistakes (difficult/easy)	8.5
5.5 Designed for all levels of users: experienced and non-experienced (never, always)	9
6. Usability and system interface	9.2
6.1 Use of colors and sounds (poor, good)	9.4
6.2 system feedback (poor, good)	9.4
6.3 System messages and reports (poor, good)	9.25
6.4 System clutter and user interface noise	8.75

5.3.3 Discussion

Based on the showcased results, the prototype demonstrated high content validity, which approves its appropriateness as a teaching tool for urethral dissection based on prostate size and prostate cancer location. Based on the feedback summarized in 5.2 and visualized in Figure 5.8, mean ratings surpassed the minimum rating threshold of $\frac{7}{10}$, except statement 3, where the rating was $\frac{5.4}{10}$. This was an expected result since expert surgeons are trained well enough to know where to dissect the urethra without the need for a tool.

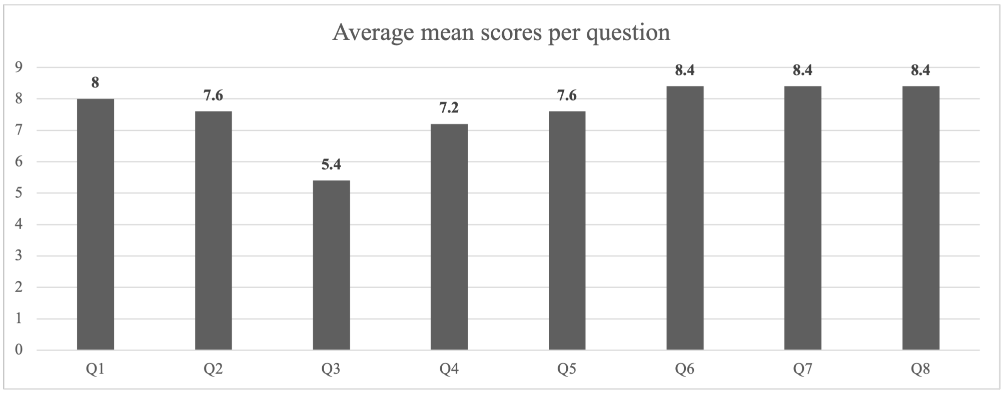


Figure 5.8: Average mean scores per question in content validity

However, during later discussions, participating surgeons mentioned that this tool might not be used to teach expert surgeons but rather to evaluate the surgeons' choice on where to dissect the urethra intraoperatively. Overall, there was no significant discrepancy between expert and non-expert surgeons in terms of ratings. Statistically, the p-value was $0.304(30.4\%)$, which means no statistical significance among the ratings of experts and non-experts. Along with the content validity, the user interaction questionnaire was satisfactory as all mean ratings were above the minimum threshold of $\frac{7}{10}$. This reflects the surgeons' satisfaction with several aspects.

As shown in Table 5.3 and Figure 5.9, six different interface aspects were evaluated: the system's overall reaction, screen, system information and terminology, learning, system capabilities, and system usability and interface. Looking at the mean ratings, question 1.5 scored the least rating with $\frac{7.6}{10}$. Although the system meets the expectation, it is crucial to refine the interface to make it more stimulating. This is important to encourage the trainee surgeons to keep using the tool for training and not feel bored while using it. Question 6.1 scored one of the highest scores among the rest. This is expected as surgeons during feedback sessions showed high impression with color coding and box illustration on the urethra. This is also reflected in statement 6 rating in content validity.

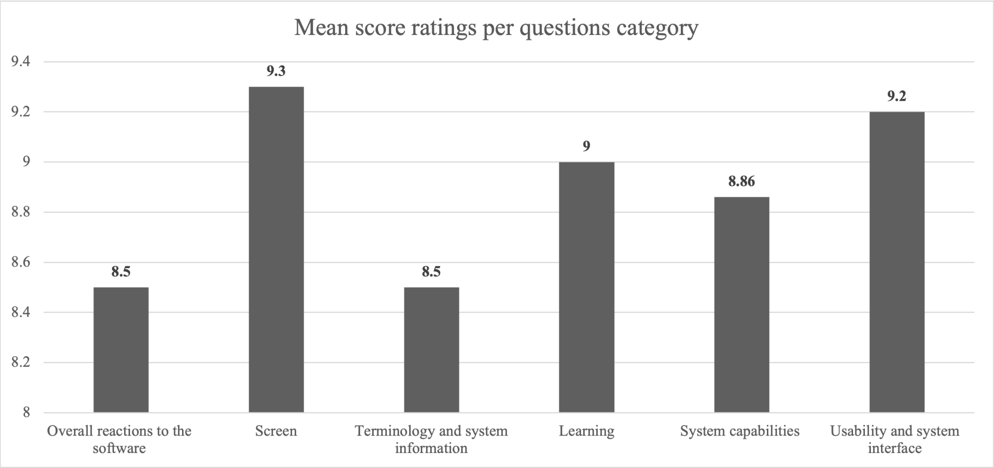


Figure 5.9: Mean score ratings per question category in QUIS

To conclude, this chapter discussed the surgeons' feedback on the tutorial mode that teaches surgeons where to dissect the urethra during a RARP procedure based on two variables: prostate size and cancer location. Feedback has proven to be satisfactory.

CHAPTER 6: CONCLUSION & FUTURE WORK

In the medical field, continuous learning and skills assessment are critical to ensure trainees and even expert surgeons to not lose their surgical skills. This can be done through various ways, including computer-based assessment training. In this thesis, we have attempted to construct a training tool for trainee surgeons to instruct them on the urethra's optimum cutting range during a robotic prostatectomy procedure. Clinically, this step is particularly crucial as an inaccurate cut may lead to leaving cancerous tissues behind and thus, jeopardizing surgical outcomes. Thesis consisted of two major parts: (1) Interface design and urethra localization using U-Net architecture and (2) Clinical validation. In the interface, trainees would train on the optimum cutting ranges in a tutorial mode. Next, and on evaluation mode, trainee surgeons would attempt to perform a cut, and based on their performance, the cut will be evaluated and scored. Our preliminary investigations have shown urologists accepted the application and perceived it well. Unfortunately, and due to the current COVID-19 situation, we are not able to perform further usability studies with surgeons.

To the best of our knowledge, this is the first training software for surgeons to train in the optimum cutting range for a prostatectomy procedure.

6.1 Limitations

Multiple limitation have been faced during the development and evaluation of this work, which may have affected the quality and execution of this thesis. The below points summarizes them:

- No benchmarked datastes, which means dataset has to be manually collected for the sake of this thesis. This task is time consuming and requires multiple prior approvals.

- Collected dataset do not have high volume of images, which may effect the deep learning results negatively.
- Dataset images have multiple issues especially issue that degrade the image quality such as smoke, presence of surgical tools, etc. Bad frames had to be taken away. Images-related issues are explained and illustrated further in Section 4.2.
- Manual annotation of images which was highly time consuming as it is highly dependant on the presence of an expert surgeon to validate the annotations.
- Unavailability of GPU-based computers, which necessitates the use of Google Collab, which may offer inconsistent results as it randomly allocates available resources and not sticking to the same resource all the time.
- Unavailability of surgeons to perform a second usability evaluation due to their tight schedule, especially during the pandemic.

6.2 Future works

For the future work, and as clinical validations have showed that this would be an important addition in the operating room, it is worth investigating incorporating this module for training purposes for new surgeons. This tool can also serve as a preoperative surgical planning tool to plan the urethra cutting location by incorporating accurate patient data and images. It is also worth investigating the potential of applying such technology into a similar scenario where a dissection region is crucial surgically.

It is also worth investigating online deep learning training model where the model is fed new labeled urethra datasets more frequently to generate even more accurate urethra segmentation until we reach an acceptable end-user accuracy level. The literature have shown that transfer learning using weights of pre-trained model enhances the results significantly as in [37] and [68] which is a path that we can investigate in order to achieve higher results. To go one level higher, we can investigate the results of deep learning approaches versus traditional image processing techniques. The next step for this project can be deployment in training rooms in resident rooms in HMC to further test the application and address limitations. We will also investigate the effectivness of using VR technology into such prototype.

Once the results are satisfactory, we can investigate the next steps of commercialization and patenting.

One publication is the current output of this work. The further anticipated publications from this projects are (1) a journal/conference paper on the usability study conducted with surgeons for the evaluation mode, and (2) a journal publication describing the urethra localization from surgical images using medical image segmentation model (U-Net).

REFERENCES

[1] A. Zemmar, A. M. Lozano, and B. J. Nelson, “The rise of robots in surgical environments during covid-19,” *Nature Machine Intelligence*, vol. 2, no. 10, pp. 566–572, Oct. 2020, ISSN: 2522-5839. DOI: 10.1038/s42256-020-00238-2. [Online]. Available: https://doi.org/10.1038/s42256-020-00238-2.

[2] L. M. Huynh and T. E. Ahlering, “Robot-assisted radical prostatectomy: A step-by-step guide,” eng, *Journal of endourology*, vol. 32, no. S1, S28–S32, May 2018, PMC6071518[pmcid], ISSN: 1557-900X. DOI: 10.1089/end.2017.0723. [Online]. Available: https://doi.org/10.1089/end.2017.0723.

[3] *Prostate cancer*. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/prostate-cancer/symptoms-causes/syc-20353087.

[4] *Annual cancer facts and figures - 2020*. [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf.

[5] *Da vinci surgical systems*. [Online]. Available: https://www.intuitive.com/en-us/products-and-services/da-vinci/systems.

[6] M. Alnazari, M. Zanaty, E. Rajih, A. El-Hakim, and K. C. Zorn, “Standardized 4-step technique of bladder neck dissection during robot-assisted radical prostatectomy,” *Investig Clin Urol*, vol. 57, no. Suppl 2, S165–S171, Dec. 2016, ISSN: 2466-0493. DOI: 10.4111/icu.2016.57.S2.S165. [Online]. Available: https://doi.org/10.4111/icu.2016.57.S2.S165.

[7] A. N. Sridhar, T. P. Briggs, J. D. Kelly, and S. Nathan, "Training in robotic surgery-an overview," eng, *Current urology reports*, vol. 18, no. 8, pp. 58–58, Aug. 2017, PMC5486586[pmcid], ISSN: 1534-6285. DOI: 10.1007/s11934-017-0710-y. [Online]. Available: https://doi.org/10.1007/s11934-017-0710-y.

[8] L. M. Huynh and T. E. Ahlering, "Robot-assisted radical prostatectomy: A step-by-step guide," eng, *Journal of endourology*, vol. 32, no. S1, S28–S32, May 2018, PMC6071518[pmcid], ISSN: 1557-900X. DOI: 10.1089/end.2017.0723. [Online]. Available: https://doi.org/10.1089/end.2017.0723.

[9] S. F. Mungovan, J. S. Sandhu, O. Akin, N. A. Smart, P. L. Graham, and M. I. Patel, "Preoperative membranous urethral length measurement and continence recovery following radical prostatectomy: A systematic review and meta-analysis," *European Urology*, vol. 71, no. 3, pp. 368–378, Mar. 2017. DOI: 10.1016/j.eururo.2016.06.023. [Online]. Available: https://doi.org/10.1016/j.eururo.2016.06.023.

[10] L. Ichim and D. Popescu, "Road detection and segmentation from aerial images using a cnn based system," in *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, 2018, pp. 1–5. DOI: 10.1109/TSP.2018.8441366.

[11] P. Wang, N. G. Cuccolo, R. Tyagi, I. Hacihaliloglu, and V. M. Patel, "Automatic real-time cnn-based neonatal brain ventricles segmentation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 716–719. DOI: 10.1109/ISBI.2018.8363674.

[12] K. Hu, C. Liu, X. Yu, J. Zhang, Y. He, and H. Zhu, “A 2.5d cancer segmentation for mri images based on u-net,” in *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, 2018, pp. 6–10. DOI: 10.1109/ICISCE.2018.00011.

[13] M. Dasgupta, O. Bandyopadhyay, and S. Chatterji, “Automated helmet detection for multiple motorcycle riders using cnn,” in *2019 IEEE Conference on Information and Communication Technology*, 2019, pp. 1–4. DOI: 10.1109/CICT48419.2019.9066191.

[14] K. Shi, H. Bao, and N. Ma, “Forward vehicle detection based on incremental learning and fast r-cnn,” in *2017 13th International Conference on Computational Intelligence and Security (CIS)*, 2017, pp. 73–76. DOI: 10.1109/CIS.2017.00024.

[15] X. Mou, X. Chen, J. Guan, B. Chen, and Y. Dong, “Marine target detection based on improved faster r-cnn for navigation radar ppi images,” in *2019 International Conference on Control, Automation and Information Sciences (ICCAIS)*, 2019, pp. 1–5. DOI: 10.1109/ICCAIS46528.2019.9074588.

[16] C. Wang and Z. Peng, “Design and implementation of an object detection system using faster r-cnn,” in *2019 International Conference on Robots Intelligent System (ICRIS)*, 2019, pp. 204–206. DOI: 10.1109/ICRIS.2019.00060.

[17] Z. Li, M. Dong, S. Wen, X. Hu, P. Zhou, and Z. Zeng, “Clu-cnns: Object detection for medical images,” *Neurocomputing*, vol. 350, pp. 53–59, 2019, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2019.04.028. [Online].

Available: https://www.sciencedirect.com/science/article/pii/S0925231219305521.

[18] Y. Liu, Z. Ma, X. Liu, S. Ma, and K. Ren, "Privacy-preserving object detection for medical images with faster r-cnn," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2019. DOI: 10.1109/TIFS.2019.2946476.

[19] C.-Y. Sun, X.-J. Hong, S. Shi, Z.-Y. Shen, H.-D. Zhang, and L.-X. Zhou, "Cascade faster r-cnn detection for vulnerable plaques in oct images," *IEEE Access*, vol. 9, pp. 24 697–24 704, 2021. DOI: 10.1109/ACCESS.2021.3056448.

[20] S. Kido, Y. Hirano, and N. Hashimoto, "Detection and classification of lung abnormalities by use of convolutional neural network (cnn) and regions with cnn features (r-cnn)," in *2018 International Workshop on Advanced Image Technology (IWAIT)*, 2018, pp. 1–4. DOI: 10.1109/IWAIT.2018.8369798.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. arXiv: 1505.04597. [Online]. Available: http://arxiv.org/abs/1505.04597.

[22] J. Walsh, N. O' Mahony, S. Campbell, A. Carvalho, L. Krpalkova, G. Velasco-Hernandez, S. Harapanahalli, and D. Riordan, "Deep learning vs. traditional computer vision," Apr. 2019, ISBN: 978-981-13-6209-5. DOI: 10.1007/978-3-030-17795-9_10.

[23] Ç. Kaymak and A. Uçar, "Semantic image segmentation for autonomous driving using fully convolutional networks," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2019, pp. 1–8. DOI: 10.1109/IDAP.2019.8875923.

[24] Z. Pan, T. Emaru, A. Ravankar, and Y. Kobayashi, *Applying semantic segmentation to autonomous cars in the snowy environment*, 2020. arXiv: 2007.12869 [cs.CV].

[25] T. L. Giang, K. B. Dang, Q. Toan Le, V. G. Nguyen, S. S. Tong, and V.-M. Pham, "U-net convolutional networks for mining land cover classification based on high-resolution uav imagery," *IEEE Access*, vol. 8, pp. 186 257–186 273, 2020. DOI: 10.1109/ACCESS.2020.3030112.

[26] G. Mutreja, S. Kumar, D. Jha, A. Singh, and R. Singh, "Identifying settlements using svm and u-net," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 1217–1220. DOI: 10.1109/IGARSS39084.2020.9324702.

[27] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie, "Anatomynet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Medical Physics*, vol. 46, no. 2, pp. 576–589, 2019. DOI: https://doi.org/10.1002/mp.13300. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13300. [Online]. Available: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13300.

[28] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Multi-label whole heart segmentation using cnns and anatomical label configurations," in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and

O. Bernard, Eds., Cham: Springer International Publishing, 2018, pp. 190–198, ISBN: 978-3-319-75541-0.

[29] S. Kletz, K. Schoeffmann, J. Benois-Pineau, and H. Husslein, "Identifying surgical instruments in laparoscopy using deep learning instance segmentation," in *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, 2019, pp. 1–6. DOI: 10.1109/CBMI.2019.8877379.

[30] C. González, L. Bravo-Sánchez, and P. Arbelaez, "Isinet: An instance-based approach for surgical instrument segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds., Cham: Springer International Publishing, 2020, pp. 595–605, ISBN: 978-3-030-59716-0.

[31] E. Yanik, X. Intes, U. Kruger, P. Yan, D. Miller, B. V. Voorst, B. Makled, J. Norfleet, and S. De, *Deep neural networks for the assessment of surgical skills: A systematic review*, 2021. arXiv: 2103.05113 [cs.CV].

[32] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Evaluating surgical skills from kinematic data using convolutional neural networks," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds., Cham: Springer International Publishing, 2018, pp. 214–221, ISBN: 978-3-030-00937-3.

[33] X. Liu, B. Zhang, A. Susarla, and R. Padman, *Youtube for patient education: A deep learning approach for understanding medical knowledge from user-generated videos*, 2018. arXiv: `1807.03179 [cs.CV]`.

[34] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C.-W. Fu, and P.-A. Heng, "Sv-rcnet: Workflow recognition from surgical videos using recurrent convolutional network," *IEEE Transactions on Medical Imaging*, vol. 37, no. 5, pp. 1114–1126, 2018. DOI: `10.1109/TMI.2017.2787657`.

[35] D. Withey and Z. J. Koles, "A review of medical image segmentation: Methods and available software," 2008.

[36] H. Fan, L. Chen, C. Feng, Z. Li, Y. Zhao, and S. Zhang, "Segmentation of corpus spongiosum from male anterior urethra ultrasound images," in *2017 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, 2017, pp. 159–165. DOI: `10.1109/ICWAPR.2017.8076682`.

[37] H. Williams, L. Cattani, W. Li, M. Tabassian, T. Vercauteren, J. Deprest, and J. D'hooge, "3d convolutional neural network for segmentation of the urethra in volumetric ultrasound of the pelvic floor," in *2019 IEEE International Ultrasonics Symposium (IUS)*, 2019, pp. 1473–1476. DOI: `10.1109/ULTSYM.2019.8925792`.

[38] Y. Qiblawey, A. Tahir, M. Chowdhury, A. Khandakar, S. Kiranyaz, T. Rahman, N. Ibtehaz, S. Mahmud, S. Al-Madeed, and F. Musharavati, *Detection and severity classification of covid-19 in ct images using deep learning*, Feb. 2021.

[39] D. Jha, S. Ali, H. D. Johansen, D. D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, *Real-time polyp detection, localisation and segmentation in colonoscopy using deep learning*, 2020. arXiv: `2011.07631 [cs.CV]`.

[40] S. Hosseinzadeh Kassani, P. Hosseinzadeh Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Automatic polyp segmentation using convolutional neural networks," in *Advances in Artificial Intelligence*, C. Goutte and X. Zhu, Eds., Cham: Springer International Publishing, 2020, pp. 290–301, ISBN: 978-3-030-47358-7.

[41] Q. Nguyen and S.-W. Lee, "Colorectal segmentation using multiple encoder-decoder network in colonoscopy images," in *2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 2018, pp. 208–211. DOI: `10.1109/AIKE.2018.00048`.

[42] S. M. Kamrul Hasan and C. A. Linte, "U-netplus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 7205–7211. DOI: `10.1109/EMBC.2019.8856791`.

[43] M. Attia, M. Hossny, S. Nahavandi, and H. Asadi, "Surgical tool segmentation using a hybrid deep cnn-rnn auto encoder-decoder," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017, pp. 3373–3378. DOI: `10.1109/SMC.2017.8123151`.

[44] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, Mar. 2019, ISSN: 2196-

1115. DOI: 10.1186/s40537-019-0192-5. [Online]. Available: https://doi.org/10.1186/s40537-019-0192-5.

[45] P. Harrison, N. Raison, T. Abe, W. Watkinson, F. Dar, B. Challacombe, H. Van Der Poel, M. S. Khan, P. Dasgupta, and K. Ahmed, "The validation of a novel robot-assisted radical prostatectomy virtual reality module," *Journal of Surgical Education*, vol. 75, no. 3, pp. 758–766, 2018, ISSN: 1931-7204. DOI: https://doi.org/10.1016/j.jsurg.2017.09.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1931720417303537.

[46] K. H. Sheetz, J. Claflin, and J. B. Dimick, "Trends in the adoption of robotic surgery for common surgical procedures," eng, *JAMA network open*, vol. 3, no. 1, e1918911–e1918911, Jan. 2020, 2758472[PII], ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2019.18911. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2019.18911.

[47] A. J. Hung, P. Zehnder, M. B. Patil, J. Cai, C. K. Ng, M. Aron, I. S. Gill, and M. M. Desai, "Face, content and construct validity of a novel robotic surgery simulator," *Journal of Urology*, vol. 186, no. 3, pp. 1019–1025, 2011. DOI: 10.1016/j.juro.2011.04.064. eprint: https://www.auajournals.org/doi/pdf/10.1016/j.juro.2011.04.064. [Online]. Available: https://www.auajournals.org/doi/abs/10.1016/j.juro.2011.04.064.

[48] L. Sessa, C. Perrenot, S. Xu, J. Hubert, L. Bresler, L. Brunaud, and M. Perez, "Face and content validity of xperience™ team trainer: Bed-side assistant training simulator for robotic surgery," *Updates in Surgery*, vol. 70, no. 1, pp. 113–119,

Mar. 2018, ISSN: 2038-3312. DOI: 10.1007/s13304-017-0509-x. [Online]. Available: https://doi.org/10.1007/s13304-017-0509-x.

[49] M. Alshuaibi, C. Perrenot, J. Hubert, and M. Perez, "Concurrent, face, content, and construct validity of the RobotiX Mentor simulator for robotic basic skills," *Int J Med Robot*, vol. 16, no. 3, e2100, Jun. 2020.

[50] P. Ramos, J. Montez, A. Tripp, C. K. Ng, I. S. Gill, and A. J. Hung, "Face, content, construct and concurrent validity of dry laboratory exercises for robotic training using a global assessment tool," *BJU Int*, vol. 113, no. 5, pp. 836–842, May 2014.

[51] C. Perrenot, M. Perez, N. Tran, J. P. Jehl, J. Felblinger, L. Bresler, and J. Hubert, "The virtual reality simulator dV-Trainer(®) is a valid assessment tool for robotic surgical skills," *Surg Endosc*, vol. 26, no. 9, pp. 2587–2593, Sep. 2012.

[52] R. G. Olsen, F. Bjerrum, L. Konge, J. V. Jepsen, N. H. Azawi, and S. H. Bube, "Validation of a Novel Simulation-Based Test in Robot-Assisted Radical Prostatectomy," *J Endourol*, Mar. 2021.

[53] D. Xiao, J. J. Jakimowicz, A. Albayrak, S. N. Buzink, S. M. Botden, and R. H. Goossens, "Face, content, and construct validity of a novel portable ergonomic simulator for basic laparoscopic skills," *Journal of Surgical Education*, vol. 71, no. 1, pp. 65–72, 2014, ISSN: 1931-7204. DOI: https://doi.org/10.1016/j.jsurg.2013.05.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1931720413001529.

[54] M. A. Mercado, R. R. Gonzalez, and B. J. Dunkin, "Proving face, content, and construct validity for a photoselective vaporization of prostate (pvp) simulator," *Journal of Urology*, vol. 189, no. 4S, e810–e810, 2013. DOI: 10.1016/j.

juro.2013.02.2392. eprint: https://www.auajournals.org/doi/pdf/10.1016/j.juro.2013.02.2392. [Online]. Available: https://www.auajournals.org/doi/abs/10.1016/j.juro.2013.02.2392.

[55] A. Alsalamah, R. Campo, V. Tanos, G. Grimbizis, Y. Van Belle, K. Hood, N. Pugh, and N. Amso, “Face and content validity of the virtual reality simulator ‘scantrainer®’,” *Gynecological Surgery*, vol. 14, no. 1, p. 18, Sep. 2017, ISSN: 1613-2084. DOI: 10.1186/s10397-017-1020-6. [Online]. Available: https://doi.org/10.1186/s10397-017-1020-6.

[56] F. Alvarez-Lopez, M. F. Maina, and F. Saigí-Rubió, “Use of a Low-Cost Portable 3D Virtual Reality Gesture-Mediated Simulator for Training and Learning Basic Psychomotor Skills in Minimally Invasive Surgery: Development and Content Validity Study,” *J Med Internet Res*, vol. 22, no. 7, e17491, Jul. 2020.

[57] E. Leijte, E. Arts, B. Witteman, J. Jakimowicz, I. De Blaauw, and S. Botden, “Construct, content and face validity of the eoSim laparoscopic simulator on advanced suturing tasks,” *Surg Endosc*, vol. 33, no. 11, pp. 3635–3643, Nov. 2019.

[58] C. Gillan, J. Papadakos, J. Brual, N. Harnett, A. Hogan, E. Milne, and M. E. Giuliani, “Impact of high-fidelity e-learning on knowledge acquisition and satisfaction in radiation oncology trainees,” *Current Oncology*, vol. 25, no. 6, pp. 533–538, 2018, ISSN: 1718-7729. DOI: 10.3747/co.25.4090. [Online]. Available: https://www.mdpi.com/1718-7729/25/6/4090.

[59] S. Hajesmaeel-Gohari and K. Bahaadinbeigy, “The most used questionnaires for evaluating telemedicine services,” *BMC Medical Informatics and Decision*

Making, vol. 21, no. 1, p. 36, Feb. 2021, ISSN: 1472-6947. DOI: 10.1186/s12911-021-01407-y. [Online]. Available: https://doi.org/10.1186/s12911-021-01407-y.

[60] N. Agarwal, S. S. Kommana, D. R. Hansberry, A. I. Kashkoush, R. M. Friedlander, and L. D. Lunsford, "Accessibility, reliability, and usability of neurosurgical resources," *J Neurosurg*, vol. 126, no. 4, pp. 1263–1268, Apr. 2017.

[61] F. Alanazi, "Evaluating the usability of the laboratory information system (lis) in coombe hospital and hail hospital," Master's Thesis, Technological University Dublin, 2015.

[62] S. Tomar, "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[63] *Rectlabel*. [Online]. Available: https://rectlabel.com/.

[64] *Computer vision annotation tool*. [Online]. Available: cvat.org.

[65] *Nvidia tesla t4*. [Online]. Available: https://www.techpowerup.com/gpu-specs/tesla-t4.c3316.

[66] *Davinci surgery community*. [Online]. Available: https://www.davincisurgerycommunity.com/.

[67] S. F. Mungovan, J. S. Sandhu, O. Akin, N. A. Smart, P. L. Graham, and M. I. Patel, "Preoperative Membranous Urethral Length Measurement and Continence Recovery Following Radical Prostatectomy: A Systematic Review and Meta-analysis," *Eur Urol*, vol. 71, no. 3, pp. 368–378, Mar. 2017.

[68] A. A. Kalinin, V. I. Iglovikov, A. Rakhlin, and A. A. Shvets, “Medical image segmentation using deep neural networks with pre-trained encoders,” in *Deep Learning Applications*, M. A. Wani, M. Kantardzic, and M. Sayed-Mouchaweh, Eds. Singapore: Springer Singapore, 2020, pp. 39–52, ISBN: 978-981-15-1816-4. DOI: 10.1007/978-981-15-1816-4_3. [Online]. Available: https://doi.org/10.1007/978-981-15-1816-4_3.

ACRONYMS

AR Augmented Reality. 1

CNN Convolutional Neural Network. vii, x, 8, 9

FCL Fully Connected Layers. 8

FDA Food and Drug Administration. 5

FN False Negative. 18

FP False Positive. 18

HMC Hamad Medical Corporation. iii, 2, 5, 24, 42, 50, 58

IoU Intersection Over Union. x, 17, 18

QUIS Questionnaire for User Interaction Satisfaction. viii, 52

RARP Robot-Assisted Radical Prostatectomy. 1, 2, 4, 6, 7, 24, 25, 42, 43

ReLU Rectified Linear Unit. 9, 10

ROI Region of Interest. 17, 23

TN True Negative. 18

TP True Positive. 17

VR Virtual Reality. 1, 58

APPENDIX A: SURGEONS FEEDBACK FOR PRELIMINARY CLINICAL VALIDATION

Q1: As an intraoperative module, if you had a tool to help you visualize/locate the optimal dissection location for urethral transection, how useful would that be? (Not useful, Somewhat useful, extremely useful)

$\frac{3}{3}$ surgeons agree that it is extremely useful for novice and intermediate surgeons to guide them through.

Q2: As an intraoperative module, if such a tool is useful, to what expertise level should it be targeting (novice, intermediate, or expert robotic surgeon)

$\frac{1}{3}$ surgeon say it is extremely useful. Even experts surgeons may use it as a "back up" to confirm their location for the urethral dissection. $\frac{1}{3}$ surgeon say somewhat useful for expert surgeons during intervention but extremely useful for trainee surgeons. Surgeon suggested to draw emphasis that this is a suggested cut and it's the responsibility of surgeon to find optimum cut. $\frac{1}{3}$ surgeon do not see that it is useful intraoperatively.

Q3: As a training module, do you think a training tool for identifying optimal cut location in urethral transection would be useful? (Not useful, Somewhat useful, extremely useful)

$\frac{3}{3}$ surgeons say it is an extremely useful tool to be able to learn in a fault-tolerant environment, generate reports, learn from mistakes and for the supervisor/trainer to monitor their progress and performance over time.

Q4: As a training module, if such a tool is useful, to what expertise level should it be targeting (novice, intermediate, or expert robotic surgeon)

$\frac{2}{3}$ surgeons believe that, for all levels of trainees, having this tool available is invaluable. Expert Surgeons are always learning and looking for ways to improve efficiency without giving up quality. If the tool made surgeons even a little bit more efficient and/or confident, it is helpful.

$\frac{1}{3}$ surgeon believe that it is useful for novice and intermediate surgeons but not for expert surgeons as they already know what they are doing and such tool will not be of major addition.