Automated Detection of Renal Masses in Contrast-Enhanced MRI using Deep Learning Methods

by

Anush Agarwal

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Applied Science
in
Engineering

Guelph, Ontario, Canada

© Anush Agarwal, August, 2021
ABSTRACT

AUTOMATED DETECTION OF RENAL MASSES IN CONTRAST-ENHANCED MRI USING DEEP LEARNING METHODS

Anush Agarwal
University of Guelph, 2021

Advisor(s):
Dr. Eranga Ukwatta

Multiparametric MRI based assessment of renal masses in the kidney has the potential to improve tumour classification accuracy and as a result improve treatment outcomes for renal cancer patients. Despite the increased use of MRI for renal mass assessment, the use of deep learning techniques to detect and classify renal masses using this modality remains unexplored. I propose a fully automated computer aided detection algorithm that identifies spatially separated renal masses in abdominal contrast-enhanced nephrographic phase MRI volumes. A cascaded series of U-Net models is used, with the first step isolating kidney boundaries with a Dice score of 91.20%. These boundaries are used to identify instances of renal masses with a mass-wise precision of 83.3% and a recall of 86.2%. The proposed algorithm can serve as a localizer for future work that incorporates multiparametric MRI data to classify tumours based on pathology.
ACKNOWLEDGEMENTS

I would like to personally thank my advisor, Dr. Eranga Ukwatta, for giving me the opportunity and resources to conduct this study. His support played a critical role throughout my graduate studies. Dr. Ukwatta’s openness to new ideas and guidance made my time as a graduate student an excellent and rewarding experience.

I would also like to thank my clinical collaborator, Dr. Nicola Schieda, for his insights which helped establish the relevance of my work. He was very patient, yet readily available to help interpret my findings from a radiologist’s perspective. Additionally, I want to thank Dr. Mohamed Elfaal, the radiology fellow that provided exceptional and prompt clinical labels as my study evolved.

I want to thank Dr. Bob Dony for taking the time to be part of my advisory committee during my graduate studies and providing me with a foundation for image processing through several courses during my undergraduate degree.

I also want to thank my fellow Guelph Medical Imaging AI lab members for their camaraderie throughout the courses, labs, and countless hours researching that we shared.

Additionally, I acknowledge Compute Canada for making available state-of-the-art computing resources which considerably accelerated the pace and scope of my work. I acknowledge NSERC for their financial support of this project. I acknowledge the School of Engineering at the University of Guelph, which sparked my interest in the field of medical imaging and opened my mind to the possibilities of technological innovation.

Finally, I want to thank my family. My grandmother’s kind words and optimism helped me keep my spirits up through all the challenges I faced. Without the constant encouragement and support of my parents, I would not be where I am today.
TABLE OF CONTENTS

Abstract ......................................................................................................................... ii

Acknowledgements ..................................................................................................... iii

Table of Contents ...................................................................................................... iv

List of Tables ........................................................................................................... vii

List of Figures ........................................................................................................... ix

List of Abbreviations ............................................................................................... xi

List of Appendices ..................................................................................................... xii

Chapter 1: Introduction ............................................................................................. 1
  1.1 Contributions ...................................................................................................... 3
  1.2 Organization of Thesis ....................................................................................... 4

Chapter 2: Background .............................................................................................. 5
  2.1 Kidney Cancer .................................................................................................... 5
  2.2 Role of Diagnostic Imaging in Kidney Cancer Diagnosis ................................. 5
  2.3 Diagnostic Imaging Techniques ......................................................................... 7
    2.3.1 Computed Tomography .............................................................................. 7
    2.3.2 Magnetic Resonance Imaging .................................................................... 7
    2.3.3 Ultrasound Imaging ............................................................................... 10
  2.4 Computer Aided Detection of Cancer .............................................................. 10
  2.5 Deep Learning ................................................................................................. 11
    2.5.1 Algorithm Overview ................................................................................ 11
    2.5.2 Optimization and Loss Functions ............................................................. 17
    2.5.3 Overfitting Prevention Techniques ............................................................ 18
    2.5.4 Transfer Learning ..................................................................................... 20
    2.5.5 Ensemble Methods ................................................................................... 22
  2.6 Blob Detection .................................................................................................. 22
2.7 Evaluation Methods ........................................................................................................... 23
  2.7.1 Segmentation ................................................................................................................. 23
  2.7.2 Detection ....................................................................................................................... 24
Chapter 3: Literature Review ................................................................................................. 26
  3.1 Non-Deep Learning Based Methods ................................................................................. 26
    3.1.1 Kidney Segmentation in MRI ....................................................................................... 26
    3.1.2 Segmentation and Detection of Renal Masses in CT ..................................................... 27
  3.2 Deep Learning Based Methods ......................................................................................... 28
    3.2.1 Kidney Segmentation and Renal Mass Detection in CT ................................................ 28
    3.2.2 Kidney Segmentation in MRI ....................................................................................... 31
Chapter 4: Development of an Algorithm for Kidney Segmentation in MRI ...................... 35
  4.1 Dataset Description ......................................................................................................... 35
    4.1.1 NG-MRI Dataset .......................................................................................................... 35
    4.1.2 CECT Dataset .............................................................................................................. 40
  4.2 Methods ........................................................................................................................... 41
    4.2.1 Preprocessing Techniques ......................................................................................... 42
    4.2.2 Model Architecture .................................................................................................... 43
    4.2.3 Transfer Learning ...................................................................................................... 48
    4.2.4 Model Development and Training ............................................................................. 49
    4.2.5 Postprocessing Techniques ....................................................................................... 50
    4.2.6 Evaluation of the Models for Segmentation Accuracy ................................................ 51
  4.3 Results ............................................................................................................................. 52
  4.4 Discussion ......................................................................................................................... 59
  4.5 Conclusion ....................................................................................................................... 62
Chapter 5: Development of an Algorithm for Renal Mass Detection .................................... 63
  5.1 Methods ........................................................................................................................... 63
    5.1.1 Preprocessing Techniques ......................................................................................... 64
    5.1.2 Model Architecture for Renal Mass Detection ............................................................ 67
    5.1.3 Detection of Spatially Separated Instances ................................................................. 70
LIST OF TABLES

Table 1: Summary of reviewed studies on kidney segmentation and renal mass detection. .... 34
Table 2: Breakdown of NG-MRI dataset by magnetic field strength and MRI manufacturer. .... 36
Table 3: Breakdown of solid renal mass pathological subtypes within the NG-MRI dataset. ..... 38
Table 4: Breakdown of the validation set by renal mass pathology. ........................................ 39
Table 5: Breakdown of test set by renal mass pathology.......................................................... 40
Table 6: Breakdown of renal mass types within the CECT dataset......................................... 41
Table 7: Kidney segmentation accuracy of 2D U-Net models with varying depth. Results are reported on the validation set.............................................................. 53
Table 8: Kidney segmentation accuracy of 2D U-Net models with varying loss functions....... 54
Table 9: CECT kidney segmentation accuracy of the U-Net model compared to the U-Net model proposed by Zabihollahy et al. in [57]................................................................. 55
Table 10: Kidney segmentation performance of the final U-Net model. ............................... 56
Table 11: Renal mass segmentation and detection accuracy on the validation set using different kidney localization techniques......................................................... 73
Table 12: Renal mass segmentation and detection accuracy based on model depth............. 75
Table 13: Renal mass segmentation and detection accuracy based on different loss functions.. 75
Table 14: CECT Renal mass segmentation and detection accuracy................................. 76
Table 15: Effect of pre-training with CECT data on renal mass segmentation and detection performance. ................................................................. 76
Table 16: Effect of localization method on healthy tissue segmentation and RM detection accuracy. ................................................................. 78
Table 17: Loss function optimization for the healthy tissue segmentation model based on RM detection recall. ................................................................. 78
Table 18: Renal mass detection accuracy based on ensemble method when using background-elimination localization.................................................................................................................................................. 79

Table 19: Renal mass detection accuracy based on ensemble method when using spatial-prior localization.................................................................................................................................................. 80

Table 20: Tumour specific detection recall and localization of the final RM detection model... 80

Table 21: Evaluation of RM detection on patient-wise. Results reported on entire NG-MRI dataset using 5-fold cross validation using model-generated kidney segmentations for localization.................................................................................................................................................. 81

Table 22: Patient-wise solid renal mass detection recall. Results reported on the entire NG-MRI dataset using 5-fold cross validation using model-generated kidney segmentations for localization.................................................................................................................................................. 81
## LIST OF FIGURES

Figure 1: Stages of renal cancer progression. ................................................................. 5

Figure 2: CM Phase DCE-MRI (left), NG Phase DCE-MRI (right). .............................. 10

Figure 3: LeNet-5 architecture. ....................................................................................... 13

Figure 4: The original U-Net architecture. ..................................................................... 16

Figure 5: Attention gates (AG) architecture, retrieved from [39]. ................................. 16

Figure 6: Axial view of the kidney in a NG-MR image (left), and its corresponding radiologist-provided segmentation where pink and yellow indicate healthy kidney tissue and red indicates clear cell RCC (right). ............................................................. 37

Figure 7: Axial view of the kidney in a CECT image (left), and its corresponding radiologist-provided segmentation where yellow indicates healthy kidney tissue, red indicates a solid renal mass, and blue indicates fluid-filled cysts (right). .......................................................... 41

Figure 8: Process flow for developing the kidney segmentation algorithm. ..................... 42

Figure 9: Architecture of a 2D U-Net model with 6 layers of skip connections. ............... 45

Figure 10: Architecture of a 3D U-Net model with 5 layers of skip connections. ............ 46

Figure 11: Architecture of a 2D Attention U-Net model with 6 layers of skip connections. 46

Figure 12: Architecture of a 3D Attention U-Net model with 5 layers of skip connections. 47

Figure 13: Kidney segmentation performance of four U-Net architectures on the NG-MRI validation set. ................................................................. 53

Figure 14: Effect of transfer learning from CECT on NG-MRI kidney segmentation DSC based on the NG-MRI training set size. ................................................................. 55

Figure 15: NG-MRI Kidney segmentation prediction axial views. ............................... 57

Figure 16: 3D views of the NG-MRI kidney segmentations. .......................................... 58

Figure 17: Receiver operator curve for the final kidney segmentation model based on the entire NG-MRI dataset using 5-fold cross validation. ................................................................. 59
Figure 18: Process for the development of the renal mass detection algorithm. .................................................. 64

Figure 19: Final process for fully autonomous renal mass detection. .............................................................. 64

Figure 20: a) Background-eliminated localization example with an image and its ground truth, b) Spatial-prior localization example with a 2-channel input of the image and kidney segmentation and its corresponding ground truth. .................................................................................. 66

Figure 21: Renal mass segmentation and detection accuracy using different deep learning architectures. .................................................................................................................. 74

Figure 22: ROC curve for the final ensemble RM segmentation model with AUC. ........................................... 82

Figure 23: a) – f) Predicted renal mass boundaries (red) are compared to ground truth renal mass boundaries (green) for patients with various renal mass pathologies. Two axial slices from each patient are shown. .................................................................................. 84

Figure 24: 3D views of renal mass segmentations with matching detections highlighted by colour. .................................................................................................................. 85
### LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG</td>
<td>Attention gate</td>
</tr>
<tr>
<td>AI</td>
<td>Artificial intelligence</td>
</tr>
<tr>
<td>AUC</td>
<td>Area under the curve</td>
</tr>
<tr>
<td>BCE</td>
<td>Binary cross entropy</td>
</tr>
<tr>
<td>CAD</td>
<td>Computer aided detection</td>
</tr>
<tr>
<td>cc-RCC</td>
<td>Clear cell - renal cell carcinoma</td>
</tr>
<tr>
<td>CD</td>
<td>Centroid distance</td>
</tr>
<tr>
<td>CECT</td>
<td>Contrast enhanced CT</td>
</tr>
<tr>
<td>CM</td>
<td>Corticomedullary</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional neural network</td>
</tr>
<tr>
<td>CT</td>
<td>Computed tomography</td>
</tr>
<tr>
<td>DCE-MRI</td>
<td>Dynamic contrast enhanced MRI</td>
</tr>
<tr>
<td>DI</td>
<td>Diagnostic imaging</td>
</tr>
<tr>
<td>DSC</td>
<td>Dice similarity coefficient</td>
</tr>
<tr>
<td>DW-MRI</td>
<td>Diffusion-weighted MRI</td>
</tr>
<tr>
<td>FN</td>
<td>False negative</td>
</tr>
<tr>
<td>FP</td>
<td>False positive</td>
</tr>
<tr>
<td>HD</td>
<td>Hausdorff distance</td>
</tr>
<tr>
<td>HU</td>
<td>Hounsfield unit</td>
</tr>
<tr>
<td>LoG</td>
<td>Laplacian of Gaussian</td>
</tr>
<tr>
<td>MoE</td>
<td>Mixture of experts</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic resonance imaging</td>
</tr>
<tr>
<td>NG</td>
<td>Nephrographic</td>
</tr>
<tr>
<td>NG-MRI</td>
<td>Nephrographic phase contrast enhanced MRI</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>PKD</td>
<td>Polycystic kidney disease</td>
</tr>
<tr>
<td>RAVD</td>
<td>Relative absolute volume difference</td>
</tr>
<tr>
<td>RCC</td>
<td>Renal cell carcinoma</td>
</tr>
<tr>
<td>ReLU</td>
<td>Rectified linear unit</td>
</tr>
<tr>
<td>RM</td>
<td>Renal mass</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operator characteristic curve</td>
</tr>
<tr>
<td>SRM</td>
<td>Small renal mass</td>
</tr>
<tr>
<td>TN</td>
<td>True negative</td>
</tr>
<tr>
<td>TP</td>
<td>True positive</td>
</tr>
</tbody>
</table>
LIST OF APPENDICES

Appendix A: Figure Reproduction Permissions ................................................................. 102
Chapter 1: Introduction

Renal cancer is the 10th most commonly diagnosed cancer in Ontario, and its incidence rate is increasing annually by 5.3% [1]. Globally, it results in 131,000 deaths and has an estimated economic burden of $1.6 billion USD. Reducing primary risk factors such as tobacco and alcohol consumption, obesity, and hypertension, as well as improving early diagnostic techniques have been identified as means of lessening the impact of kidney cancer on society [2]. Early diagnosis of kidney cancer due to symptoms such as flank pain, haematuria (blood in the urine), and a palpable mass in the abdomen are rare because these symptoms generally only present in late stages where treatment is more challenging [3]. Instead, over 50% of diagnoses occur incidentally when an abdominal CT, MRI, or ultrasound is performed for an unrelated purpose and a mass is observed and investigated.

Renal masses include fluid-filled cysts or solid tumours and may be benign or malignant. Once incidentally detected, the pathology and staging of a renal mass must be completed. The primary medical tool for this has historically been contrast enhanced CT (CECT) imaging, where an intravenously injected contrast agent accumulates in kidney tissue [4]. While CECT imaging is effective at detecting renal masses and differentiating cysts from solid tumours, it struggles at differentiating benign solid tumours from malignant, with one study reporting CECT to have a specificity of diagnosing RCC to be 27.7% [5]. In situations where the malignancy of a renal mass is indeterminate after CECT assessment, clinicians have historically been inclined to surgically remove the renal mass and potentially the entire affected organ in a procedure called a nephrectomy. In cases where the renal mass is in fact benign, this practice leads to overtreatment because the mass may have been managed by active surveillance. Overtreating a benign renal mass exposes patients to the risks of nephrectomies such as compromised kidney function, hospital acquired infections, and surgical complications. In fact, overtreatment has been identified as the primary reason mortality rates for kidney cancer patients has not improved in recent years, despite increased early incidental detection rates [6]. Tissue biopsy is another diagnostic technique which can help determine renal mass malignancy, however it is not used often despite having the highest classification ability of all diagnostic techniques. The risks and costs of renal tissue biopsy, which is also an invasive procedure, often outweighs its benefits [7].
Multi-parametric dynamic contrast enhanced MRI (DCE-MRI) based renal mass assessment has recently gained clinical interest for its improved ability to determine the malignancy of renal masses compared to CECT, with one study reporting an RCC specificity of 68.1% (vs 27.7% with CECT) [5]. With greater certainty of renal mass pathology, clinicians may opt for less invasive treatment options more often, leading to a reduction in overtreatment related mortality. MRI is less invasive and has virtually no risk of complications compared to tissue biopsy while being more accurate than CECT. It therefore may see greater clinical adoption for renal mass assessment than renal mass biopsy.

Despite increased interest in using MRI for renal mass assessment, to the best of my knowledge there have been no attempts to develop an autonomous computer aided detection (CAD) tool to identify renal masses or classify renal masses using MRI. CAD tools and classification tools for renal masses have been extensively studied using CECT images, primarily because CECT imaging has seen the most use historically based on its ubiquitous availability, speed, and cost. Novel techniques that attempt renal mass classification using multiparametric MRI may benefit from its increased soft tissue contrast to achieve higher accuracy than those using CECT images. A CAD tool that uses MRI data is also relevant to clinical practices as MRI is increasingly used for renal mass assessment. Such a tool may improve early detection rates by reducing inter-operator variability and improving detection recall, potentially leading to better outcomes for patients. There are challenges associated with developing an effective renal mass CAD tool in MRI. Since MRI is used less often for renal mass assessment, there is less data available for it than CECT. Furthermore, while MRI has better soft tissue contrast than CECT, it has poorer geometric accuracy and spatial resolution.

The objective of the work presented in this thesis is to develop a fully autonomous renal mass detection algorithm using abdominal DCE-MRI based on deep learning methods. The terms ‘autonomous’ and ‘automated’ in this thesis are not related with automated machine learning (Auto-ML) [8], where model architectures and hyperparameter settings are determined without human input, but rather the end result of the detection task. The motivation behind this work is to establish a baseline for renal mass detection capability using MR images. The algorithm proposed in this thesis is also intended to serve as a localizer for future renal mass classification algorithms that can leverage the strengths of the MRI modality to identify the malignancy and pathology of detected masses. To achieve the objective of autonomously detecting renal masses,
a fully automated kidney segmentation algorithm is also proposed. Furthermore, to address limitations related to the small size of the available MRI dataset, transfer learning using CECT data is also considered. Final evaluation of the model is completed by assessing its ability to detect individual spatially separated instances of renal masses within MR images.

1.1 Contributions

My work contributes to the fields of computer aided detection and medical image segmentation through its application in a novel area and its unique technical implementation. This section highlights specific contributions and the novelty of the work presented in this thesis.

- In Chapter 4, a deep learning algorithm that autonomously predicts kidney boundaries in DCE-MRI volumes is proposed. This work is unique because the segmentation of renal mass containing kidneys in MRI volumes has not been attempted. All MRI volumes used in this study were verified to contain renal masses through biopsy or surgery. Additionally, the kidney segmentation accuracy of the proposed model is evaluated thoroughly based on DSC score, precision, recall, Hausdorff distance, and average volume difference.

- In Chapter 5, a deep learning algorithm that autonomously detects renal masses in DCE-MRI volumes is proposed. The detection of renal masses in MRI volumes has not been attempted before, making this a valuable contribution. Compared to similar work that detects renal masses in CECT, the proposed implementation is unique. A mixture-of-experts style ensemble is attempted where renal masses and healthy kidney tissue are segmented by different models. Additionally, an asymmetric loss function called \( F_\beta \)-BCE loss is implemented to address dataset imbalance and prioritize recall during training. Finally, the proposed model is capable of detecting spatially separated masses in 3D. The accuracy of this model is also evaluated in 3D on a mass-wise basis.

- For both Chapter 4 and 5, cross-modality transfer learning using CECT data is attempted to improve results compared to independently using the MRI data. This has not been attempted before for kidney segmentation or renal mass detection.
1.2 Organization of Thesis

The organization of this thesis is as follows:

Chapter 1: Introduction

The objective of this thesis is introduced, with an emphasis on the contributions made by the presented work.

Chapter 2: Background

Topics related to the presented work are described. An overview of renal cancer and the role diagnostic imaging techniques play in detecting and treating it are discussed. The origins of the deep learning technique used in the presented work is described, including an overview of evaluation methods for segmentation and detection.

Chapter 3: Literature Review

A review of prior work related to fully autonomous kidney segmentation and renal mass segmentation and detection is completed. The strengths and weaknesses of previous approaches is discussed, and the novelty of the proposed work is put into context.

Chapter 4: Development of an Algorithm for Kidney Segmentation in MRI

An autonomous kidney segmentation algorithm using whole abdominal MRI volumes is proposed. The methodology is described, and evaluation of segmentation accuracy is completed.

Chapter 5: Development of an Algorithm for Renal Mass Detection

A deep learning based renal mass detection algorithm is proposed. This work builds on the autonomous kidney segmentation algorithm proposed in Chapter 4. The proposed model is described in detail and evaluated based on its renal mass detection capability.

Chapter 6: Conclusion and Recommendations

Findings from the work presented in this thesis are summarized. Recommendations for next steps and future work are made based on the insights gained from this study.
Chapter 2: Background

2.1 Kidney Cancer

Renal cell carcinoma (RCC) can be divided into pathological subtypes based on its appearance in histopathology slides. Clear cell RCC, which accounts for 85% of all malignant tumours, is characterised by the pale and clear appearance of its cells. Papillary RCC is the second most common type at 10%, characterised by long finger-like projections at the cellular level. The third most common is chromophobe RCC making up approximately 5% of masses, characterised by its larger cell size [3]. Amongst benign solid masses, the most common sub-pathology is angiomyolipoma, characterised by the presence of vasculature and muscle tissue. The second most common benign sub-pathology is oncocytoma, characterised by epithelial cells with abnormal amounts of mitochondria.

The stage of a tumour as shown in Figure 1, will influence patient outcomes and treatment plans. Generally, the larger the diameter of a tumour, the more likely it is to be cancerous. One study reported the risk of renal mass malignancy increasing by 16% with every 1 cm increase in diameter [9]. Small renal masses (SRM) with a diameter less than 4cm, are benign 20% of the time and are the focus of this thesis [10].

Figure 1: Stages of renal cancer progression. Retrieved from [11].

2.2 Role of Diagnostic Imaging in Kidney Cancer Diagnosis

The clinical gold standard for investigating renal masses with diagnostic imaging (DI) is contrast-enhanced computed tomography (CECT) imaging, although MRI and ultrasound may
be used situationally [3]. DI is able to effectively differentiate between tumorous (solid) and cystic (fluid-filled) renal masses and helps identify the stage of a tumour.

Distinguishing malignant from benign SRMs in early stages with DI alone remains to be a challenge, and often leads to overtreatment [12]. CECT imaging is unreliable at making this distinction due to their similar appearance [4]. MRI is increasingly being recommended for helping make this distinction due to its improved soft-tissue contrast, although its use is historically uncommon [5]. Another diagnostic tool is renal mass biopsy, where cores of tissue from suspected regions are surgically extracted and analyzed in a lab. While this method is around 92% accurate at determining pathology, the current standard of care only recommends biopsy for patients at high risk of adverse surgical outcomes where biopsy may justify the prevention of surgery [10][13]. Otherwise, the risks of biopsy such as bleeding, tumour seeding, and insufficient tumour sampling, are considered too severe to justify its use [14].

The primary treatment of RCC is surgically through a nephrectomy [3]. This procedure involves resecting either the entire organ or, in the case of partial nephrectomy, resecting the affected region of the organ. Nephrectomy is recommended in the majority of cases, except when the patient is of high risk of surgical complications or the mass is confirmed to be benign [3][13]. In practice, the latter scenario does not happen often because of the low diagnostic ability of CT imaging and the lack of widespread use of tissue biopsy or MRI based confirmation. The overuse of nephrectomy in the treatment of potentially benign masses has been identified as the root cause behind why mortality rate has not decreased for kidney cancer patients despite an increase in early detection rates [4]. Through widespread DI use, SRMs are being identified earlier and more often, resulting in administration of potentially life saving treatment earlier on in the progression of the disease. Unfortunately, overtreatment with nephrectomy concurrently increased the incidence of non-oncologic deaths due to surgical complications such as reduced kidney function and infection, negating the benefits of early detection.

Considering the risks of overtreatment, a better approach to improving outcomes for kidney cancer patients using DI is to not only detect masses early, but also determine their malignancy with greater confidence. This information will help clinicians make better informed treatment decisions. The current approach focused on CECT imaging is deficient in this regard, with one study reporting a diagnostic accuracy of 79.41% for classifying RCC malignancy [15]. Increased
use of biopsy may help address this issue due to their higher accuracy, but Urologists’ apprehension of the procedure due to its invasive nature and risk of complications such as hemorrhage are a significant barrier to widespread use [4]. MR imaging has the potential to meet this clinical need by providing superior malignancy classification ability to CT, with one study reporting a diagnostic accuracy of 88.1% for classifying RCC, while being less invasive to a patient than biopsy [15].

2.3 Diagnostic Imaging Techniques

2.3.1 Computed Tomography

Computed tomography (CT) imaging measures the attenuation of x-ray photons by tissue to produce a 3D image [16]. Tissue types within a voxel can be identified by their level of attenuation, with the standard unit of measurement for attenuation being the Hounsfield unit (HU). CT imaging benefits from this standardization of intensities, as well as having relatively quicker imaging times, lower maintenance costs, and greater geometric accuracy than MRI imaging, often making it the modality of choice. A significant drawback of CT is its use of ionizing radiation in imaging which can lead to a permanent increase in lifetime risk of cancer. A typical abdominal CT scan delivers 10-20 mSv of radiation, attributable to a lifetime risk of developing a solid cancer of 1 in 1000 [17].

Contrast enhanced computed tomography (CECT) is the clinical standard imaging technique for renal masses [18]. The standard imaging protocol for renal CECT involves intravenous injection of an iodine-based contrast agent. Areas that accumulate more of the iodine based agent appear hyperintense on the image due to increased attenuation of the x-ray signal. Images are acquired temporally at different stages of contrast permeation. A scan is acquired before injection representing the pre-contrast baseline, followed by a corticomedullary (CM) phase scan taking place 30-40s after injection, and finally a nephrographic (NG) phase scan after another 80s [18]. The phase corresponds to where the contrast agent is concentrated in the organ. NG phase CECT images are the most common and clinically useful for renal mass detection [4].

2.3.2 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) applies a magnetic field to detect the nuclear magnetic resonance (NMR) signal of hydrogen nuclei within tissue to produce an image [19]. The
frequency of the NMR signal, referred to as the Larmor frequency, \( \omega \), is influenced by the
\[ \text{gyromagnetic ratio of hydrogen atoms, } \gamma_H, \text{ and the strength of the applied magnetic field, } B_0, \text{ based on the Larmor equation below.} \]
\[ \omega = \gamma B_0 \]   \hspace{1cm} (1)

The gradient coil in an MRI machine varies \( B_0 \) along the \( x, y, \) and \( z \) axes of the imaging plane by
pre-defined amounts to provide spatial encoding to the resultant \( \omega \) signals [19]. When hydrogen
atoms are aligned with \( B_0 \), they are at equilibrium. However, when another local magnetic field,
\( B_1 \) is applied to the atoms in the transverse direction to \( B_0 \), the proton poles are tipped out of
alignment. Once \( B_1 \) is removed, the poles return to alignment with \( B_0 \) because it is their lowest
energy state, at which point the NMR signal can be measured along the transverse plane. Local
variations in tissue type, density, as well as the occurrence of diffusion and flow can change the
rate at which poles return to their lowest energy state, forming the basis for the differences in
contrast seen in MR images.

Energy is released by atoms as they return to equilibrium in two ways, longitudinal spin-lattice
relaxation (\( T_1 \)), and transverse spin-spin (\( T_2 \)) relaxation [19]. \( T_1 \) relaxation occurs as protons
release the excess energy from \( B_1 \) into the surroundings. It is strongly influenced by the ability of
the surrounding tissue to absorb and dissipate heat released at the Larmor frequency. Fat tissue is
known to have a significantly lower \( T_1 \) relaxation time than water, allowing the two to be
differentiated. \( T_2 \) relaxation occurs as the local magnetic fields induced by individual protons
interfere with each other, causing their transverse magnetization to de-phase. This is influenced
by the rate of motion of the particles, with those moving closer to the Larmor frequency having
the quickest \( T_2 \) relaxation, such as fat and water-based solutions.

\[ S(\text{TR, TE}) = \left( 1 - e^{-\frac{\text{TR}}{T_1}} \right) M_0 e\frac{\text{TE}}{T_2} \]   \hspace{1cm} (2)

The NMR signal measured in the transverse plane is influenced by both relaxation types as seen
in Eq. (2). To control which relaxation has the greatest impact on image contrast and produce a
\( T_1 \)-weighted or \( T_2 \)-weighted image, the Echo Time (TE) and Repetition Time (TR) are adjusted
[19]. The repetition time controls how much longitudinal magnetization recovers before the \( B_1 \)
field is re-applied. If \( TR \) is chosen such that \( T_1 \) is still in the transient phase, protons with a
quicker \( T_1 \) time will have a larger longitudinal magnetization and as a result, a larger transverse
magnetization upon repetition. As per Eq. (2), if TE is concurrently chosen to be close to 0, the influence of T2 relaxation is largely removed producing a T1-weighted image. Conversely, choosing a large TR causes its corresponding term in the equation to simplify to 1 for all tissues irrespective of their T1 properties, leading to a T2-weighted image. Note that the $M_0$ term in the equation above refers to the Boltzmann Magnetization of a proton. This captures the net magnetization of a proton based on its quantum properties, temperature, and the applied magnetic field strength $B_0$.

In the context of Diagnostic Imaging (DI) the difference in T1 and T2 times of healthy and diseased tissue may be very small, leading to low contrast [20]. Dynamic contrast-enhanced MRI (DCE-MRI) is a technique where a gadolinium based agent is injected into the patient intravenously. Gadolinium has the effect of locally reducing T1 relaxation time. This makes areas with higher concentration of the contrast agent appear brighter in T1-weighted images. Contrast naturally tends to accumulate in renal tumours at varying rates compared to healthy kidney tissue due to the localized hypervascularity of tumours, making T1-weighted contrast imaging particularly useful for tumour detection [20]. Within contrast imaging techniques, the delay between contrast injection and image acquisition will determine which region the contrast accumulates in. While values vary, a kidney MRI taken approximately 35s after contrast administration is referred to as a corticomedullary (CM) phase image, and an MRI taken later around the 120s mark is referred to as a nephrographic (NG) phase image [21]. The NG phase of contrast enhancement is valued for renal mass detection due to its sensitivity for lesion detection [20]. As per the visualization of both phases in Figure 2, the contrast spreads more evenly throughout the kidney in NG compared to CM where it is concentrated around the renal cortex.

The true value of MRI however lies in its multiparametric use where each sequence can highlight different pathologies, leading to a higher diagnostic accuracy than CECT imaging [5]. As per [21], a typical kidney MRI imaging protocol for renal mass investigation will include:

- A T2-weighted sequence.
- Chemical shift imaging with In-Phase and Opposed-Phase components.
- Fat suppressed T1-weighted sequences before and after gadolinium injection.

A clinician will draw conclusions by combining information from each sequence. For example, T2-weighted imaging is effective at differentiating RCC from AMLs, chemical shift imaging is effective at classifying cc-RCC, and contrast-enhanced MRI is effective at lesion identification.
Image acquisition times vary widely, with a typical 5-sequence multiparametric MRI taking 20-60 minutes [22].

![Figure 2: CM Phase DCE-MRI (left), NG Phase DCE-MRI (right).](image)

**2.3.3 Ultrasound Imaging**

Renal masses can be found incidentally during ultrasound imaging of the abdomen. Ultrasound imaging measures the acoustic impedance of tissue to produce an image [23]. The amount of fluid within an imaged tissue will have a strong impact on its intensity on the final image. Highly liquid regions such as cysts have lower impedance than connective tissue such as bone, resulting in hypointense regions. Naturally, this property makes ultrasound imaging effective at differentiating cysts from solid renal masses [24]. Furthermore, ultrasound imaging may be used during partial nephrectomies to assess the extent of tumour spread and identify metastasis within the vasculature surrounding the kidney. The strength of this modality lies in its low cost and portability, although it is not commonly used to stage and diagnose renal masses due to significant flaws. The ultrasound signal is susceptible to noise and artifacts from interfering intestinal gasses and unfavorable body compositions [25]. Additionally, ultrasound imaging is often unable to distinguish between the renal fascia and the fat layer that surrounds the kidney, making it difficult to stage a tumour. Due to the highlighted issues, ultrasound imaging use in the management of renal masses is primarily limited to incidental detection. Once a mass is identified by ultrasound it is investigated further with CT or MR imaging for diagnosis.

**2.4 Computer Aided Detection of Cancer**

Computer aided detection (CAD) systems have been used to augment radiologist diagnosis workflow when working with diagnostic images. The primary goal of CAD systems are to
reduce observational oversights, where a radiologist misses a visually apparent area resulting in a false negative [26]. It should be noted that CAD has also been used in literature to abbreviate Computer aided diagnosis, where an algorithm provides a diagnosis of a mass based on pathology [26]. The focus of this thesis is on detection, so studies related to diagnosis are considered out of scope. A more conventional method to reduce oversights is double-reading, a form of peer-review where two radiologists independently assess the same case and compare results. While double-reading is effective at improving sensitivity and diagnostic accuracy in many applications, it effectively doubles the workload of radiologists making it an unsustainable solution [27]. In contrast, CAD systems deploy algorithms to highlight regions of interest that may correspond to clinically relevant masses missed by an expert reader. Since these algorithms rely on computational resources, they are more efficient than double-reading while still improving sensitivity. No CAD tool is currently in clinical use for renal cancer, although there are examples in other areas such as breast cancer screening. For example, an artificial intelligence (AI) based tool called cmAssist is being commercially developed for clinical use to highlight regions of interest in diagnostic mammograms leading to improved detection rates [28]. CAD systems such as cmAssist rely on radiologists to make the final diagnosis of any regions of interest. Their main benefit lies in reducing the risk of reading radiologists missing potentially relevant regions. Most of the latest developments in CAD systems use deep learning-based AI techniques to segment and isolate regions of interest. For example, fully automated CAD methods for detecting prostate tumours have recently been proposed [29][30].

2.5 Deep Learning

2.5.1 Algorithm Overview

Artificial intelligence-based methods have increasingly become the mainstay of medical image segmentation research. The main draw of AI-based methods is their ability to automatically learn model parameters that result in optimal performance. As stated by Yann LeCun et al. in [31], natural data in any form has such a high degree of variability and richness that designing an accurate pattern recognition algorithm by hand becomes nearly impossible. Modern machine learning algorithms are the result of innovations in algorithm design, optimization techniques, and computer hardware. To understand how a generalized deep learning algorithm can
effectively segment a region of interest in a medical image, a brief summary of neural network based machine learning algorithms is presented.

The core structure behind modern image segmentation architectures such as U-Net discussed later in this chapter is the perceptron [32]. This biologically inspired unit accepts inputs from a set of features \( x \), each of which are multiplied by a tunable weight parameter \( w \), and a tunable bias parameter \( b \). The summation of all inputs is scaled using a non-linear function, referred to as an activation \( \sigma \), to produce an output as per Eq. (3). The decision boundary of a single perceptron can be visualized as a linear hyperplane, where its weights define the hyperplane vector orientation, and its bias defines the hyperplane offset. Perceptrons can be arranged in hierarchal patterns where the output from one set of units is used as inputs for the next set of perceptrons, forming an artificial neural network. In the case of multi-layer neural networks, non-linear decision boundaries can be formed with increased classification ability.

\[
\sigma(w^T x + b) = \begin{cases} 
1, & \text{if } w^T x + b > 0 \\
0, & \text{if } w^T x + b \leq 0
\end{cases} \tag{3}
\]

The conventional neural network is dense, it requires all input features to be connected to all neurons in a subsequent layer. This requirement becomes computationally prohibitive when working with a large set of features such as medical images, where each pixel is an input. Furthermore, finding optimal weight and bias parameters for a given task is a non-trivial task due to the co-dependent nature of each layer of a neural network.

The next relevant innovation in machine learning occurred when the convolutional neural network (CNN) was introduced in 1998 by Yann LeCun et al. [31]. This algorithm uses convolutional kernels as feature extractors and treats each kernel element as a model parameter to be optimized through gradient descent. These feature extractors benefit from the constraint that their inputs must be presented simultaneously and be proximally located. Requiring location proximity encouraged neural network weights to perform complex spatial operations such as edge detection and de-noising. Furthermore, by forcing each set of weights within a kernel to be applied to every spatial region in an input image through convolution, the CNN discourages the duplication of feature extractors, reducing the computational cost compared to dense neural networks. A key feature of CNNs is their translation invariance. This property stems from the subsampling operation seen in Figure 3 which takes the feature maps outputted by a
convolutional kernel and down samples it to half its size. When downsampling, the resultant feature map is invariant to translations within its receptive field. As a result, subsequent convolution operations have a larger receptive field relative to the original image despite their kernel size being static in the height and width direction. While earlier layers of the CNN extract simple features such as edges and corners, later layers can leverage these low-level features to detect complex features such as faces, animals, trees, etc. [33]. In the context of image classification, typical CNN architectures terminate with densely connected neural layers that output the probability of an image belonging to a class based on one-hot encoding.

The work of [31] showed CNN models moderately outperforming methods such as SVM and KNN on handwritten character recognition tasks while using significantly less system memory. In the following years, CNN models increased in size to tackle increasingly complex multi-label classification tasks as a result of increases in computational capacity and the availability of large, annotated datasets. A breakthrough study was published in 2012 by Krizhevsky et al., leading to widespread research interest in CNN models [34]. Krizhevsky et al. achieved record-breaking image classification performance with a CNN with 60 million parameters on the ImageNet dataset, a collection of 1.2 million images with 1000 distinct classes. While the general model architecture of [34] resembles Figure 3, the model size and amount of training data increased considerably. This work served as a proof of concept of the scalability of CNNs with regards to the complexity of a classification task, resulting in its widespread use in classification and segmentation tasks today. Additionally, it highlighted the importance of optimizing model performance by preventing over-fitting and performing hyperparameter tuning.

In contrast to classification of an entire image, semantic segmentation refers to the process of producing pixel-wise classifications within an input. Early efforts to apply CNNs to medical
image segmentation tasks attempted to treat semantic segmentation as a conventional classification problem. The work of [35] used a sliding-window approach to show non-overlapping patches of electron microscopy images of neural tissue to a CNN model similar to [34], except the output provides a binary classification of whether the pixel at the center of the patch belongs to the foreground. While this method was record-setting for its time, it required making a compromise between localization accuracy and contextual accuracy. For example, if the authors made the size of the sliding-window too small it may miss a large edge denoting the boundary of a neuron. Making the window larger would help provide this context, but it would also make the output lower resolution resulting in a loss of fine-grain accuracy. Advancements in CNNs became more relevant to medical image segmentation when the U-Net architecture was proposed [36].

U-Net is a fully convolutional model, similar to [37], except there is a decoder pathway which includes convolutional layers with a similar number of feature maps as the encoder pathway, resulting in a symmetrical model [36]. The encoder path is visible on left of Figure 4. It consists of two convolution layers and a down-sampling operation arranged in repeating blocks. Each block extracts features using convolutions and then encodes the features into a lower resolution space using max-pooling. Max-pooling is a form of down-sampling where the maximum value in a receptive field is used as the downsampled value as opposed to averaging all values in the receptive field. It has empirically been shown to improve CNN performance compared to downsampling by averaging [38]. Notably, the number of filters used in a convolution layer doubles at every depth level of encoder block in the model. The intuition behind this being that filters in earlier layers of CNNs perform low-level tasks such as de-noising and edge detection which are common to many high level features. The subsequent layers are more diverse as they perform specialized high-level feature extraction, and as such benefit from a larger number of filters.

Following the encoder path, the decoder or expanding path replaces the max-pooling operation with an upconvolution operation. This operation performs a transpose convolution to double the size of each feature map. Since the weights of the convolution transpose kernel are learnable parameters, it is possible to achieve superior upscaling compared to standard interpolation methods [37]. The final up-convolution produces feature maps of the same size as the input image, assuming convolution is performed with padding. While the other convolution layers had
a kernel size of 3x3, the final convolution layer has a kernel size of 1x1 and, for a binary semantic segmentation problem, a single filter. The primary purpose of the final convolution is to combine all the channels of the feature map into one and apply a sigmoid activation, resulting in the output segmentation. All other convolutional layers use a Rectified Linear Unit (ReLU) activation function, which has been shown to reduce training time within deep learning models [34].

The reason U-Net is more effective at fine-grain semantic segmentation than the model in [35] is its use of skip connections between layers of identical resolution, represented by the copy and crop function in Figure 4. Due to the CNN’s translation invariance, there is an unavoidable loss of localized information when an image is downsamples. Skip connections attempt to reintroduce this lost information by concatenating feature maps that were created before downsampling. U-Nets are considered the gold-standard for medical image segmentation, often producing near radiologist level accuracy. Additionally, the U-Net is fully convolutional, resulting in a relatively smaller number of parameters compared to CNNs that include dense layers [36]. This results in reduced computational load during training and requires less training data to prevent over-fitting.

Since the inception of U-Net, several variations have been proposed. In [39], the 3D U-Net was proposed. Each operation in the original U-Net is effectively replaced by its 3D counterpart allowing for segmentation of volumetric data. This method was motivated by a desire to use sparsely annotated medical images as a way to generate volumetric segmentations. Additionally, a 3D feature extractor may be more effective as it can capture elements not apparent in 2D slices of images. There are several disadvantages to the 3D approach. A 3D U-Net of the same size as a 2D U-Net has significantly more parameters to learn due to the additional dimension, resulting in the need for larger annotated datasets which is a challenge to acquire due to the resource-intensive nature of voxel-level labelling. Furthermore, a single patient volume will only count as one training sample to a 3D model, whereas for a 2D model each slice can be used as a separate data point. Finally, 3D models also need specialized computing hardware as they occupy more system memory and often take longer to train [40]. Despite these drawbacks, the U-Net and its variations are commonly used in various medical image segmentation tasks due to their strong performance [41]–[44].
Another variation of interest to the U-Net architecture is the Attention U-Net [45]. This model uses attention gates (AG) as seen in Figure 5 to apply soft region proposals to skip connection feature maps in a U-Net model. Soft region proposals form a resampler that scales each skip connection activation by a value between $[0,1]$ indicating its saliency. The coefficients of the attention resampler are determined using upsampled features from the previous convolution layer. AGs reduce false positives by suppressing activations that are outside the region of interest suggested by low resolution feature maps [45].
2.5.2 Optimization and Loss Functions

The standard approach for optimizing deep learning model parameters involves using backpropagation of a loss function relative to model parameters, first applied to a neural network by Rumelhart et al. in 1986 [46]. This process is commonly referred to as gradient descent. To execute gradient descent, a loss value that measures model performance must be calculated. In the case of a simple binary classification problem, a function such as binary cross entropy (BCE) seen in Eq. (4) may be used. This function forms the objective, or cost function, of the model with a zero loss indicating ideal performance for a given sample set. The gradient of the loss function, \( \nabla L \), is its partial derivative relative to each of the model parameters. It forms a vector that indicates the magnitude and direction by which to adjust each parameter to maximize change to the loss function. By finding average gradients across multiple training examples, this method aims to autonomously suggest effective model parameters that are well generalized based on a global improvement in the loss function. This process can be repeated iteratively, with each iteration producing weights that further minimize the cost function. The goal of gradient descent is to eventually produce parameters resulting in convergence of the cost function to its global minimum, a point of theoretically optimum performance. In practical use, convergence to the global minimum is unlikely due to the presence of local minima, however despite this fact gradient descent has been shown to be effective in parameter optimization [46]. Conventional gradient descent optimization requires the loss to be calculated across all training examples before updating parameters. A stochastic form of gradient descent where parameters are updated after only observing a small subset, or batches, of all examples is more commonly used due to improvements in computation speed [31]. The Adam optimizer further modifies stochastic gradient learning by setting unique learning rates for each parameter which factors in the mean and variance of previously calculated gradients, resulting in improved convergence time [47].

\[
BCE(w_0, w_1, ..., w_n) = -\frac{1}{N} \sum_{n=1}^{N} G_n \log(P_n) + (1 - G_n) \log(1 - P_n)
\]  

(4)

Other loss functions have been proposed to train deep learning models. Cross entropy loss, described in Eq. (4), is a common choice for medical image segmentation tasks and is used in the original U-Net [36]. An alternative choice is the use of the Dice Similarity Coefficient (DSC) as a loss function [48]. DSC is commonly used as an evaluation metric for segmentation, so
incorporating it into a loss function enables the model to directly train towards better evaluation. Since DSC outputs a range from 0 to 1, where 1 is a perfect segmentation, it can be converted to a loss function as per Eq. (5). Here, \( P \) and \( G \) represent the predicted and ground truth values for each pixel \( N \), with \( \epsilon \) being a small value added to prevent a divide by zero error. Using Dice loss enables the model to directly optimize the metric that will be used to judge its performance. It also responds well to imbalanced segmentation data where there are few foreground pixels in a test image.

\[
\text{Dice Loss} = 1 - \frac{2 \times \sum_{n=1}^{N} P_n G_n + \epsilon}{\sum_{n=1}^{N} P_n + G_n + \epsilon}
\]  

An extension to Dice loss called the asymmetric similarity loss layer was proposed in [49], which can be used to train a model that unequally prioritizes precision and recall. The DSC in a binary segmentation task can be thought of as the harmonic mean of pixel-wise precision and recall where equal priority is given to optimizing precision and recall. In asymmetric similarity loss, referred to as \( F_\beta \) loss in this thesis due to its similarity to the \( F_\beta \)-score, unequal prioritization can be given to precision or recall by modifying a \( \beta \) parameter as per Eq. (6). The relative component terms \( G_n \setminus P_n \) and \( P_n \setminus G_n \) correspond to the false negatives and false positives, respectively. This feature is relevant to the medical field since incorrectly classifying a positive diagnosis often causes more harm than incorrectly classifying a negative one.

\[
F_\beta \text{ Loss} = 1 - \frac{\sum_{n=1}^{N} P_n G_n + \epsilon}{\sum_{n=1}^{N} P_n G_n + \frac{\beta^2}{1 + \beta^2} G_n \setminus P_n + \frac{1}{1 + \beta^2} P_n \setminus G_n + \epsilon}
\]  

2.5.3 Overfitting Prevention Techniques

A significant challenge in developing deep learning algorithms is their tendency to perform worse when tested on unseen real world data. A major cause of this is overfitting, where model parameters incorrectly fit noise within the training data, resulting in poor model generalization [50]. CNNs are prone to overfitting due to their large number of parameters and due to the automatic selection of model parameters. The risk of overfitting can be reduced by regularizing parameters. Regularization penalizes the model for increases in complexity caused by regularizing parameter values or creating an over-reliance on a sub-set of features. The general goal of regularization can be explained in the context of gradient descent as preventing the model from
converging to an undesirable local minima, which can prevent the model from achieving optimal performance [46].

2.5.3.1 Data Augmentation
A large and diverse training dataset is essential to training a well generalized machine learning model. Increases in the training set size make it a more representative sample of the true population and make overfitting less likely. In cases where the sample size is small, data augmentation can be used to synthetically introduce variation to the training data by applying geometric or intensity based transformations to original images [34]. Ronneberger et al. proposed that data augmentation using elastic deformations is well suited to medical imaging tasks because it realistically models common tissue variations and deformations [36]. Elastic deformations have since seen widespread adoption in the field, with many leading solutions in the KiTS19 kidney and tumour segmentation challenge implementing it to reduce overfitting [51]. Other types of data augmentations commonly used in training deep learning models include random rotations, scaling, mirroring, Gaussian noise, and contrast changes [52].

2.5.3.2 Dropout
Dropout is a common method of regularization used during training to prevent the co-adaption of feature extractors [53]. Co-adaption occurs when extracted features only provide useful information when paired with other extracted features. These pairings may be specific to over-represented patterns in the training set, reducing generalization. With dropout, every activation on a feature map has a non-zero chance of being set to zero during training. This encourages the extraction of features that have limited dependence on other features to be useful.

2.5.3.3 Training with a Validation Set
Validation sets play an important role in quantifying overfitting since this data is not used for updating model parameters in the training phase. After every epoch of training, the loss of the model can be calculated on the validation set along with the training set. When the training loss continues to improve but the validation loss stabilizes, it is an indication that the model is learning to detect patterns which are over-represented in the training set. This information can be incorporated at train time by stopping training when improvements to the validation loss cease despite improvements to the training loss, a process referred to as early stopping [54].
Additionally, consistent use of the validation set prevents overfitting to the test set during the development of deep learning models. Overfitting the test set occurs when practitioners select hyperparameters based on test loss. Despite the test data not being directly used to update model parameters, its use in choosing hyperparameters makes it a less reliable measure of real world performance. This is why it is common for organizers of grand challenges such as KiTS19 to retain their test sets and instead require contestants to submit their model weights for assessment [51]. A similar approach can be used in development by using the validation set to determine hyperparameter performance and only determining test set performance after the model is finalized.

### 2.5.3.4 K-Fold Cross Validation

Overfitting occurs when patterns that are over-represented in the training set are learned [53]. However, the same logic may be applied to the set of data used to determine model performance. Validation and test sets may also over-represent some patterns compared to their true distributions, making their assessment biased. Additionally, validation and test sets are significantly smaller than the training set. Increasing the size of the evaluation sets would reduce the risk of biases, however this would require reducing the size of the training set which would also increase the risk of overfitting.

K-fold cross validation attempts to reduce overfitting while circumventing the trade off between validation, test, and training set size [55]. Instead of training a single model, multiple models are trained in parallel with each one reserving a different set of data for validation. For example, in 5-Fold cross validation all samples are divided into five sets. Each set is treated as the validation set for one of the models, with the remaining samples used for training. Validation results are then found by averaging performance across all folds. The main advantage of this approach is that validation performance metrics are more generalized. A drawback of cross validation is that it can significantly increase development time and resource consumption, since each model must effectively be trained K number of times more.

### 2.5.4 Transfer Learning

Medical datasets for supervised learning problems are often small because annotating them is a resource intensive task that requires domain expertise. Transfer learning methods allow models to leverage learning in a different but related task, referred to as a source task, to improve
performance in a target task [56]. In its simplest form, transfer learning is implemented by pre-training an entire model on a source task [57]. After pre-training, the weights are fine-tuned on the target task. In the context of kidney and renal mass segmentation, there is significantly more CECT imaging data available than DCE-MRI data because CECT is the clinical standard [4]. Despite fundamental differences in how the CT and MR signals are acquired, the high-level task of segmenting a kidney or renal mass boundary shares some similarity between the two. This similarity originates from both signals correlating with the morphological shape of the kidney and their common use of intravenous contrast agents leading to a stronger signal in the same anatomical region. Compared to the conventional approach where model weights are randomly initialized, in transfer learning the weights initialize in a state that is closer to their optimal values, potentially leading to improved performance.

Transfer learning across medical imaging modalities is an emerging area in research with studies showing an improvement in training time and the need for less data in the target task [57]. No studies were found attempting cross-modality transfer learning for kidney or renal mass segmentation. The work of [56] attempted transfer learning for liver segmentation from MRI to CT images and found that any improvement in organ segmentation accuracy was negligible but reported an improvement to the loss convergence speed. Another study by [58] found that using transfer learning enables an MRI liver segmentation model to generalize to a CT liver segmentation model with as few as 10 volumes of CT images, but did not test whether there was an improvement in performance compared to a randomly initialized model when using the full target dataset. The effect of cross-modality transfer learning is highly task specific, with cases that have the smallest training sets benefiting the most.

The potential benefit of transfer learning is allowing models to adapt to variations of the same task with limited data while leveraging previously known information. In the context of renal mass classification, this can have benefits beyond cross-modality training. For example, transfer learning may be used to adapt the model proposed in this thesis to variations in MRI scans from different healthcare systems that may use unique imaging protocols. Transfer learning may also be used to transfer knowledge across segmentation tasks [56].
2.5.5 Ensemble Methods

The performance of a deep learning algorithm can be improved by training multiple variations of models and combining their outputs, a process known as ensemble methods [59]. Ensemble methods have a history of being used in leading medical image segmentation models including the nnU-Net framework which was the top submission in the KiTS19 challenge [52][51]. A common approach seen in recent works is to creating an ensemble is to train models of different architectures on the same task [60][61]. An alternate approach to creating an ensemble is the mixture of experts method (MoE) proposed in [62] where a task is simplified into subtasks with an independent model specializing in each. The outputs of each model in a MoE approach is combined together using a gating function. The use of multiple models in ensemble methods increases the representational capacity of the algorithm, potentially leading to a boost in performance. In the case of MoE ensembles, the increase in representational capacity is the result of each component being trained on a different objective, leading to unique features being extracted. Additionally, the pooling of information from multiple models mitigates the risk of biases within individual models, reducing overfitting risk.

2.6 Blob Detection

The task of blob detection refers to isolating regions with coherent properties in an image. In the context of SRM detection, each spatially separated mass can be considered to be blob. Deep learning based medical CAD methods commonly use segmentation models to propose semantic maps of blobs. These semantic maps are binary and spatially separated masses must be individually identified to provide instance detection. The work presented in this thesis uses Laplacian of Gaussian (LoG) based blob detection as a processing tool to extract instances of renal masses from a binary semantic segmentation. LoG is a blob detection technique which provides scale-invariant blob detection [63]. The process of LoG blob extraction can be broken up into three stages. First, an image is convolved with a Gaussian kernel to eliminate noise. Next a Laplace operator is applied to the output of the Gaussian kernel, where blobs create strong responses. Finally, this process is repeated with Gaussian kernels of varying sizes to capture blobs at different scales. LoG blob extraction returns the coordinates of points that produce large responses in the output of the Laplacian and their corresponding scale. By controlling what kernel sizes are used for the Gaussian filter, the sizes of the detected blobs can be controlled.
This is relevant to renal mass detection as masses can occupy a large range of sizes and gives the ability to order blobs by their size. Coordinates from LoG detectors are used as initialization points for a connected-component search which defines the pixel-wise boundary of each detected instance within the binary segmentation.

The LoG detector is used for its simplicity in how it detects blobs of various sizes. Its notable limitation is its tendency to be over-sensitive to blob edges, or one directional changes in intensity, and its geometric instability relative to local affine transformations [64]. While other blob extraction algorithms have been proposed to deal with the over-sensitivity and geometric instability of LoG, these limitations are not significant for the use case of LoG in this thesis. Test images are never augmented in any way, so location instability relative to transformations is not applicable. Oversensitivity is also not an issue since duplicate blobs which have overlap in the binary segmentation are rejected using region growing within the binary segmentation mask.

2.7 Evaluation Methods

2.7.1 Segmentation

When computing segmentation performance, results in this thesis were determined voxel-wise for each patient and averaged across all patients in the set. The primary method for assessing segmentation performance is the Dice similarity coefficient (DSC) seen in Eq. (7). While the DSC calculation produces values between 0 and 1, scores are often reported as a percentage from 0 to 100. For uniformity, DSC numbers in this thesis are presented as a percentage. DSC penalizes false positives and false negatives equally as both would reduce the intersection component relative to the union component. Furthermore, the ubiquitous use of Dice as a performance measure in medical segmentation enables results to be placed in context to other studies. Other segmentation measures used are the maximum Hausdorff Distance (HD) and the Relative Absolute Volume Difference (RAVD) as per Eq. (8) and Eq. (9) respectively. The receiver-operator curve (ROC) for the final kidney and renal mass boundary prediction models are plotted to observe their segmentation quality. The area under the curve (AUC) is also measured for the final models to quantify the results of the ROC curve.
\[ DSC = \frac{2P \cap G}{P \cup G} \times 100 \]  

(7)

\[ HD(P, G) = \max_{p \in P} \left\{ \min_{g \in G} \{d(p, g)\} \right\} \]  

(8)

\[ RAVD = \left| \frac{\sum_{n=1}^{N} P_n}{\sum_{n=1}^{N} G_n} - 1 \right| \]  

(9)

2.7.2 Detection

Metrics for the detection of renal masses were computed mass-wise for this thesis. For this approach, each renal mass in the dataset is considered to be an independent test point. As opposed to segmentation, there is no consistently used metric for renal mass detection in recently published works. The work of [61] used slice-wise overlap for detection, and the work of [65] used kidney-wise volume based comparison. Neither of these approaches are able to provide the number of renal masses and their locations within a kidney. A slice-wise approach cannot determine if masses detected in separate slices are for the same mass or a different mass. The kidney-wise approach in [65] only determines if a given kidney contains a renal mass, not how many are there. Additionally, both methods are susceptible to their own localization errors as discussed in Chapter 3.2 Deep Learning Based Methods.

The fully automated detection algorithm proposed in this work aims to provide instance detection of masses in addition to semantic segmentation of each instance, similar to the task of panoptic segmentation but for a binary case [66]. To assess the detection performance of this approach, I proposed using mass-wise voxel overlap. This detection method can be considered an extension of the slice-wise pixel overlap used in [61] to 3D with instance detection. It involves comparing the voxel locations in each predicted mass instance in a kidney to the voxel locations of all ground truth masses in the same kidney. If there is a non-zero overlap, the predicted mass is a True Positive (TP). If there is no overlap for a given predicted mass, it is
counted as a False Positive (FP). Finally, if there are masses in the ground truth that do not overlap with any predicted mass, they are counted as False Negatives (FN). A summary of this approach can be seen in the pseudocode in Eq. (10). Note that this metric does not contain a definition for True Negatives (TN), but that value is not required to determine the precision and recall of the detection model as per Eq. (11) and Eq. (12). Precision and recall scores based on this detection metric are more robust to localization errors. The focus on mass-wise assessment also results in a more meaningful detection measure than a patient-wise approach because it penalizes for missing renal masses in a kidney with multiple masses. Apart from precision and recall scores, detection performance is measured mass-wise by finding the average Euclidean Centroid Distance (CD) between a TP and its ground truth. This provides insight as to the localization accuracy of each detected mass since the non-zero voxel-overlap method does not assess the amount of overlap.

\begin{align}
\text{LET} & \quad P = \text{Predicted Masses}, \; G = \text{Ground Truth Masses}, \\
& \quad TP = \text{True Positives}, \; FP = \text{False Positives}, \\
& \quad FN = \text{False Negatives} \\
\text{FOR each mass } g \text{ in } G: & \quad \text{FOR each mass } p \text{ in } P: \\
& \quad \quad \text{IF } p \text{ intersects with } g \text{ AND } p \text{ is not already a TP:} \\
& \quad \quad \quad p = TP \\
& \quad \quad \text{BREAK} \\
& \quad \text{IF no mass in } P \text{ matches } g: \\
& \quad \quad g = FN \\
\text{FOR each mass } p \text{ in } P: & \quad \text{IF } p \text{ is not a TP:} \\
& \quad \quad p = FP
\end{align}

\begin{align}
\text{Precision} & = \frac{TP}{TP + FP} \\
\text{Recall} & = \frac{TP}{TP + FN}
\end{align}
Chapter 3: Literature Review

3.1 Non-Deep Learning Based Methods

3.1.1 Kidney Segmentation in MRI

Kidney segmentation in MRI has been used in clinical practice to diagnose chronic kidney diseases including polycystic kidney disease (PKD), renal artery stenosis, and renal parenchymal disease [67]. In these cases, the volume of the kidney is expected to change due to the disease process. DCE-MRI is often used to measure the volume of the renal medulla specifically, since this region contains the bulk of the nephrons. Segmentation methods in this area can be divided into manual and semi-automated categories. The first category involves a subject matter expert performing voxel-level labelling of kidneys in an imaging volume. While this method is still partially in use today to generate ground truth references for training fully automated data-driven approaches, it is often impractical in real world use. Manual segmentation of one kidney on an MRI can take 15-30 minutes and suffers from high inter- and intra-operator variability [67].

Semi-automated methods require initialization and often post-processing by a subject matter expert. While such approaches are faster than manual segmentation, their requirement of manual input makes them less desirable and still subject to inter- and intra-operator variability. Recently, [68] proposed a semi-automated automated kidney segmentation method from pre-localized Diffusion-Weighted MR Images (DW-MRI), achieving a DSC score of 95%. This method combined shape priors with level-set segmentation based on extracted intensity features and fourth-order spatial features to define the kidney boundary. It is considered to be a semi-automated approach because MR images needed to be localized to the kidney region before kidney segmentation was completed. Another study by Mignani et al. in [69] used region growing to determine the volume of kidneys with PKD in 30 MRI volumes and reported an average volume difference of -0.6 ± 9.6%. This region growing method is considered semi-automatic because it required the placement of a seed point within the kidney organ to begin segmentation. Many studies on kidney segmentation in MR focused on organs with PKD, which is relatable to our dataset in the sense that we include cysts as renal masses for detection purposes. However, our dataset also includes solid renal tumours which could pose unique segmentation challenges. The focus of older non-deep learning based studies such as [69] was...
often on volume estimation not segmentation, and as such did not use DSC to evaluate performance making them difficult to compare to deep learning based studies [67]. The vast majority of new studies performing kidney segmentation use deep learning based methods. For example, in recent segmentation grand challenges involving the kidney such as KiTS19 and CHAOS, the leading solutions all used feature learning based using neural network architectures rather than conventional methods [51][70].

3.1.2 Segmentation and Detection of Renal Masses in CT

At the time of writing, there is no literature on CAD systems for solid renal mass segmentation and detection in MRI. The lack of literature in renal mass segmentation in MR imaging may be attributed to the historical preference towards CT imaging for renal mass investigation. Instead, works that used non-deep learning based methods on renal mass segmentation and detection in CT imaging are discussed in this section. These works use manually defined feature-engineering based methods, and as a result the applicability of such techniques to MRI is unknown. Significant differences in the imaging modalities may make feature extractors that work in CT unusable in MRI. In contrast, neural network based techniques are known to be highly generalizable, making deep learning algorithms that perform renal mass segmentation in CT more insightful to the task in MRI than non-deep learning based methods.

Semi-automated approaches for renal mass segmentation in CT have been proposed. Kim et al. in [71] selected seed points for a region growing algorithm for renal tumour segmentation in 12 CT images. Their study measured detection on a slice-wise basis for slices containing lesions and reported an 85% sensitivity. In [72], texture and intensity based features are extracted from the kidneys of 100 CT images and inputted into an SVM classifier which determined if the patients kidney contained renal cell cancer, reporting a patient-wise sensitivity and specificity of 84% and 92% respectively. Their detection method does not provide localized information since features are aggregated for each patient.

Amongst feature-engineering based fully automated techniques, radiomics is often used to extract features that can be used by a classification algorithm. Radiomics involves extracting quantitative spatial and intensity based features from images that are not apparent to the human eye but are known to be unique to the target classification [73]. In the area of renal mass
analysis, its use has so far been focused on lesion classification, leaving its applicability to the segmentation and detection of renal masses unexplored [74].

### 3.2 Deep Learning Based Methods

#### 3.2.1 Kidney Segmentation and Renal Mass Detection in CT

While there are deep learning based studies attempting kidney segmentation in MRI, they do not compare to the wealth of literature published on the analogous task in CT. Discussion in this area is therefore initiated by reviewing the latest deep learning based works in CT. The U-Net architecture and its variations have taken a dominant role in recent studies on medical image segmentation. The following is a review of the latest work on renal organ and tumour segmentation and detection using AI-based solutions in CT. Recently, the 2019 Kidney and Kidney Tumour Segmentation challenge (KiTS19) aimed to create a globally recognized benchmark dataset for the renal organ and tumour segmentation task in CT imaging [51]. It contained 300 CT volumes with radiologist verified kidney and tumour segmentations. Performance was measured using DSC on a test set of 90 images, with the highest scoring team achieving a DSC of 97.4% and 85.1% on the kidney and tumour segmentation task respectively. Reviewing KiTS19 submissions is valuable since all submissions were trained and tested on equal grounds, enabling a fair comparison.

The top scoring team at KiTS19 used a 3D U-Net with leaky ReLU activations and five encoder blocks, starting with 30 feature maps and doubling to a maximum of 320 feature maps [75]. The model was chosen based on a self-configuring architecture framework called nnU-Net proposed by authors from the same team as the top submission [52]. The nnU-Net framework was created to address the highly task-dependant performance of different configurations of the U-Net and the cumbersome process of optimizing co-dependant model parameters. It automates the selection of pre-processing strategies, network topology, and ensemble usage. Notably, the nnU-Net paper and their KiTS19 submission highlighted that model configuration optimization is more important for performance than architectural changes.

The work of [76] was also based on the nnU-Net framework but used a three-stage cascaded U-Net architecture. The first stage localized the kidney region from CECT images using a low resolution coarse segmentation. The second stage used the full resolution CECT images to
perform high resolution segmentation of the kidney and tumour. Finally, the last stage refines the tumour segmentation further. Their approach reported a DSC score of 96.7% and 84.5% on KiTS19. Comparing to [75], the inferior results of the coarse to fine segmentation approach indicate it may not make a valuable addition to the algorithm.

The third place submission in KiTS19 proposed the VB-Net architecture [51]. VB-Net is similar to a 3D U-Net with the most significant modification being the introduction of bottlenecks within each encoder block. The conventional U-Net encoder block has multiple convolutional layers with the same number of filters. In VB-Net encoder blocks, the central convolutional layer contains fewer filters. The authors claim the feature map level encoding reduces model size and training time without significantly compromising on performance. Their submission reported a DSC score of 97.3% and 83.2% on kidney and tumour segmentation, respectively.

All the top scoring submissions used a form of the U-Net deep learning architecture, despite there being no explicit constraint requiring the use of a deep learning based solution. This provides empirical evidence of the dominance of deep learning based methods, and the U-Net architecture, in the field of medical image segmentation tasks. The top submissions also favored the use of 3D U-Nets although did not make an explicit comparison to a 2D counterpart and often reporting a long training time. For example, the top approach reported training for a single model taking 5 days using a high performance GPU [75]. Additionally, the authors noted that the standard U-Net architecture continues to produce competitive results, with modifications such as residual units and other architectural changes only marginally improving results [51].

In terms of renal mass segmentation, the main limitation of the KiTS19 challenge was its lack of focus on instance segmentation or detection, with performance measured only based on semantic segmentation to simplify the annotation task. While accurate semantic segmentation of renal tumours may be able to serve as an independent detection tool, its failure to differentiate between instances of renal masses could reduce its usability as inputs to a renal mass pathology classification tool. Specifically, a classification algorithm ingesting semantic segmentations of tumours would fail in cases where masses of different pathologies are present on the same kidney. Furthermore, if the goal of renal mass assessment is detection, then using a semantic segmentation based metric is not ideal to determine performance.
The work of Lin et al. in [65] also attempted kidney segmentation in CT images and additionally attempted the detection of renal masses. Their study used a private dataset containing 441 volumes with radiologist annotated kidney and tumour segmentations. Using a series of cascaded 3D U-Net architectures, the study reported a kidney segmentation score of 97.3%, a tumour segmentation score of 84.4%, and a cyst segmentation score of 53.6% based on DSC. These results were accomplished by using an initial 3D U-Net model to perform a rough kidney segmentation which localized the kidney region, a second 3D U-Net for fine resolution kidney segmentation, and then a final model for multi-class tumour and cyst segmentation. Tumour and cyst detection was performed by identifying the largest volume of interest in the predicted segmentation and defining a volume threshold where a volume above the threshold was counted as a positive detection. Using a volume threshold of 363.3 mm$^2$, they reported a detection sensitivity of 97.4% and a specificity of 96.8%. This detection mechanism is a limitation of the study due to its use of a volume based threshold and the authors comment that missed renal tumours could be less than the threshold. While the detection results indicate high accuracy, using a volume based threshold is not ideal since it is known that tumour volume can be lower than the threshold [9]. The importance of accurate classification increases as tumour size decreases since smaller tumours are more likely to be benign [9]. Furthermore, this metric does not consider localization errors when comparing volumes of predictions to ground truths. A false positive mass of the right volume but in an incorrect location compared to the ground truth would still be counted as a true positive. A final limitation of this detection method is that analyzing only the largest volume in a segmentation excludes any possibility of detecting multiple masses in the same kidney.

Another study attempted kidney segmentation and renal mass detection using the same private dataset of CT images used in this thesis [61]. This study by Zabihollahy et al. used cascaded U-Net models with an ensemble approach. They performed kidney segmentation using a 5-block 2D U-Net architecture similar to the original architecture proposed by Ronneberger et al. [36]. A kidney segmentation DSC score of 95.79% ± 5.16% was reported on a test set of 125 volumes, and a score of 96.25% ± 3.37% on the KiTS19 dataset. The kidney segmentations were used to localize input images before feeding them to a renal mass segmentation model. Renal mass segmentation was performed using two U-Nets, one with a depth of 3 blocks and another with a depth of 4, and the final prediction mask was the addition of both outputs. A DSC score of
88.65% ± 7.31% was reported for renal masses on the private dataset of CT images. Notably their study found 2D U-Nets to deliver similar performance to 3D U-Nets while using fewer computing resources. While all segmentation metrics were reported in 3D, this study reported detection sensitivity and specificity on the basis of slice-wise overlap as follows. Any slice that had non-zero overlap between the predicted and ground truth renal mass segmentation was a true positive detection. If a slice contains renal mass in the ground truth and not in the prediction, it is labelled a false negative. Conversely, if a slice has a predicted renal mass segmentation where there is nothing in the ground truth it is a false positive. The slice-wise overlap metric of detection is an improvement over the work of [65] since it attempts to address localization error and accommodates detections of all sizes. However, it is still limited in the sense that 2D information is used as a proxy for the detection of a 3D object. For example, this metric would penalize for missing a tumour at its axial boundaries where only a few pixels are present, even if the tumour is successfully detected in all other axial slices. Additionally, it is susceptible to in-plane ambiguity, a different type of localization error. This error can be explained by considering a hypothetical scenario where two renal masses are present in a kidney separated in the x and y plane, but not the z (axial) plane, and one mass is missed completely by the segmentation algorithm while the other is perfectly segmented. Since one tumour has a non-zero overlap with the ground truth, the slices would still be counted as true positives. This makes the detection evaluation process flawed, since no penalty is given for the missed renal mass. A complete detection assessment metric should accommodate the presence of multiple renal masses in-plane. The extensive work in CT for kidney segmentation, renal mass segmentation, and renal mass detection paves the way for similar work in MRI that provides more diagnostic value due to increased soft tissue contrast. Zabihollahy et al. additionally proposed algorithms for classifying renal masses in CT by differentiating between cysts and solid masses in [77] and differentiating between malignant and benign solid masses in [78]. These two works present a strong future direction to expand upon the renal mass detection work in MRI presented in this thesis.

3.2.2 Kidney Segmentation in MRI

The previously highlighted studies represent the latest work in kidney segmentation and renal mass detection in CT imaging. This area benefitted greatly from the KiTS19 competition. It provided an open-source baseline against which model performance can be compared fairly.
Since CT imaging is the most common modality for renal mass assessment, the studies also benefited from relatively large pools of data to assess performance. In contrast, there is no established baseline for kidney and renal mass segmentation in MRI. Some studies have attempted kidney segmentation in MRI using U-Net as highlighted below, but no studies have been found that attempt renal mass segmentation or detection in MRI. Despite this, the success of U-Net based algorithms in the CT modality indicates the viability of such algorithms for complex segmentation and detection tasks. The architectures, image processing and data engineering techniques used in previous studies can help inform best practices for MRI segmentation tasks.

With regards to kidney segmentation in MRI, a study in 2018 reported a DSC score of 91.4% on a dataset of 30 pediatric T₁ weighted DCE-MRIs [79]. The DSC of 91.4% is only for non-renal mass containing kidneys as the focus of their study was not on tumour segmentation. Furthermore, their focus on pediatric MR images makes their work less applicable to the segmentation of kidneys containing cancer, a disease that primarily affects older populations with a peak incidence rate being seen in the 65-69 age group [2]. Their study used cascaded 3D U-Nets, with the first localizing the kidney region and the next performing fine-resolution segmentation. Each model performed multi-class segmentation by simultaneously distinguishing between the left kidney, right kidney, and the background.

Another study by [40] reported a DSC score of 95.6% on a dataset of 64 neck to knee T₁ weighted MRI through 8-fold cross validation. Notably, their ground truth segmentations did not include the renal hilum region which differs visually from renal parenchymal tissue. This region is included in the radiologist provided segmentations of kidneys in the dataset used in this thesis. Furthermore, their study was not focused on renal masses and there was no explicit inclusion of kidneys with tumours or cysts in their dataset. The segmentation model used was a 2.5D U-Net where the encoder weights were pre-trained on the ImageNet database. This is a variation of the 2D U-Net described by [36] where instead of inputting a single axial slice, the adjacent slices are also included forming a 3-channel image. The 2.5D model only performs inference on the center slice. The authors claim that 2.5D inputs provide additional context to the model without the overhead computational cost of a full 3D model, although do not provide comparative results against a standard 2D U-Net. Their choice of not using a 3D U-Net was motivated by the drastic
increase in training time. A 3D U-Net equivalent to their 2.5 U-Net reportedly took one day to train compared to 30 minutes for their proposed approach.

Additionally, a grand challenge called Combined Healthy Abdominal Organ Segmentation (CHAOS) included MRI kidney segmentation as a part of its effort to find an effective multi-organ segmentation model [70]. It included 40 MRI volumes each of T₁ In-Phase and T₁ Opposed-Phase sequences with registration and including segmentations of the liver, kidney, and spleen. The top scoring model on the multi-organ segmentation task was the nnU-Net with a DSC score of 94.6%, although kidney specific performance not reported. The lack of kidney specific performance reporting and the exclusion of tumour containing organs makes CHAOS results less relevant to this thesis. Although, the competitive performance of the U-Net and the nnU-Net specifically mirror results seen in KiTS19. All reviewed studies performing kidney segmentation in MRI were able achieve DSC scores in the 90% range, despite their small dataset size. None of the reviewed studies attempted renal mass segmentation in MRI which is attempted in this thesis.
Table 1: Summary of reviewed studies on kidney segmentation and renal mass detection.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Type of Method</th>
<th>Number of Patients / Modality</th>
<th>Kidney Detection Metric</th>
<th>Renal Mass Detection Metric</th>
<th>Renal Mass Detection Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isensee et al. (2019) [75]</td>
<td>3D U-Net</td>
<td>300 CECT Images</td>
<td>97.4%</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Hou et al. (2020) [76]</td>
<td>Cascaded 3D U-Nets</td>
<td>300 CECT Images</td>
<td>96.7%</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Mu et al. (2021) [51]</td>
<td>Cascaded 3D U-Nets</td>
<td>300 CECT Images</td>
<td>97.3%</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Tuncer et al. (2018) [72]</td>
<td>SVM</td>
<td>100 CT Images</td>
<td>N/A</td>
<td>Patient-wise Binary Thresholding</td>
<td>84% Sensitivity, 92% Specificity</td>
</tr>
<tr>
<td>Lin et al. (2021) [65]</td>
<td>Cascaded 3D U-Nets</td>
<td>441 CT Images</td>
<td>97.3%</td>
<td>Patient-wise Volume Thresholding</td>
<td>97.4% Sensitivity, 96.8% Specificity</td>
</tr>
<tr>
<td>Zabihollahy et al. (2020) [61]</td>
<td>2D U-Net Ensemble</td>
<td>315 CECT Images</td>
<td>95.79%</td>
<td>Slice-wise Pixel Overlap</td>
<td>87.4% Sensitivity, 81.8% Specificity</td>
</tr>
<tr>
<td>Shehata et al. (2018) [68]</td>
<td>Level Set with Shape Priors</td>
<td>64 DW-MRI</td>
<td>95%*</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Haghghi et al. (2018) [79]</td>
<td>Cascaded 3D U-Nets</td>
<td>30 Pediatric DCE-MRI</td>
<td>91.4%*</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Langner et al. (2020) [40]</td>
<td>2.5D U-Net</td>
<td>64 T1-Weighted MRI</td>
<td>95.6%*</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Isensee et al. (2021) [70]</td>
<td>3D U-Net Ensemble</td>
<td>40 T1 IP- and OP- MRI</td>
<td>94.6%*</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*Kidneys with tumours were not explicitly included in these kidney segmentation datasets.*
Chapter 4: Development of an Algorithm for Kidney Segmentation in MRI

The purpose of this chapter is to propose a fully automated algorithm for segmenting the kidney boundary using MR images. This algorithm forms the first phase of the renal mass detection algorithm by isolating the kidney region. Isolating the kidney region will simplify the renal mass detection task by limiting its search space. This approach of cascaded models has been successful in analogous tasks of kidney segmentation and renal mass detection in CECT images in [61] and [65]. Furthermore, adult kidney segmentation in NG phase DCE-MRI has not been reported in literature, making this study a valuable contribution to the field.

4.1 Dataset Description

4.1.1 NG-MRI Dataset

The data for this project was provided by The Ottawa Hospital (TOH). The use of this data has been approved by the research ethics board of TOH and approved for secondary use by the research ethics board of the University of Guelph. A total of 118 anonymized MR images were retrospectively collected between the dates of January 1st, 2015 and December 31st, 2017 and provided as DICOM files. These images included patients that underwent partial nephrectomy, total nephrectomy, or tissue biopsy for a renal mass of axial diameter less than 4cm and had an MRI performed prior to histological confirmation of pathology. The MRI volumes were collected by TOH from multiple healthcare facilities in the greater Ottawa region which used machines from different manufacturers and varied image acquisition protocols as seen in Table 2. All MR images used in this thesis were post-contrast enhanced T1 weighted fat-suppressed gradient recalled echo sequences in the axial plane collected ~100-120s post contrast injection, corresponding to the nephrographic phase. The NG-MRI sequence was chosen for this study since it is the most useful sequence for the detection task clinically [14].
Table 2: Breakdown of NG-MRI dataset by magnetic field strength and MRI manufacturer.

<table>
<thead>
<tr>
<th>Dataset Property</th>
<th>Number of Volumes</th>
<th>Median Axial Image Size</th>
<th>Median Axial Resolution</th>
<th>Median out-of-plane Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnetic Field Strength</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5 T</td>
<td>75</td>
<td>224x320</td>
<td>1.25mm</td>
<td>3.0mm</td>
</tr>
<tr>
<td>3.0 T</td>
<td>43</td>
<td>512x512</td>
<td>0.8594mm</td>
<td>4mm</td>
</tr>
<tr>
<td>MRI Manufacturer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Siemens</td>
<td>74</td>
<td>210x320</td>
<td>1.25mm</td>
<td>2.25mm</td>
</tr>
<tr>
<td>GE Medical Systems</td>
<td>40</td>
<td>512x512</td>
<td>0.8594mm</td>
<td>4.0mm</td>
</tr>
<tr>
<td>Philips Medical Systems</td>
<td>4</td>
<td>320x320</td>
<td>1.377mm</td>
<td>4.45mm</td>
</tr>
<tr>
<td>ALL NG-MRI</td>
<td>118</td>
<td>224x320</td>
<td>1.1718mm</td>
<td>3mm</td>
</tr>
</tbody>
</table>
4.1.1.1 Manual Annotations of the Data

An abdominal radiology fellow identified renal masses within the MR images and corroborated findings against the surgical and pathology reports of the patient. This ensured that renal mass pathology labels provided by TOH are free from inter-observer variability of the radiologists. Histopathological reporting of renal masses is considered to be the most accurate form of pathology assessment available, although it too has a diagnostic error rate of approximately 8% [10]. After identifying masses, the radiologist performed pixel-wise manual segmentation of the kidney and renal mass boundaries on axial slices of all 118 volumes in the nephrographic phase of the MRI (NG-MRI) using ITK-SNAP [80]. The segmentations were provided as NIfTI files with pixel-wise integer labels of the data.

If the radiologist identified additional solid masses without histological confirmation, they were segmented but their pathology was left unknown. The renal mass segmentations included cysts and tumours, although any mass less than 1cm was ignored since in these cases the accuracy of diagnostic imaging is inadequate for clinical assessment [4]. In total, the NG-MRI dataset contains 210 renal masses composed of 131 solid masses and 79 cysts. The solid renal masses included pathology labels indicating their subtype with the exception of 2 cases which had metastatic cancer from another organ, and 9 masses identified during segmentation without subtype labels. A complete breakdown of renal masses and their properties can be found in Table 3.
Table 3: Breakdown of solid renal mass pathological subtypes within the NG-MRI dataset.

<table>
<thead>
<tr>
<th>Renal Mass Type</th>
<th>Renal Mass Subtype</th>
<th>Number of Masses</th>
<th>Mean Volume (cm(^3))</th>
<th>Mean Diameter (cm)</th>
<th>Proportion of All Masses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solid Renal Mass</td>
<td>Clear Cell RCC</td>
<td>63</td>
<td>13.99 ± 10.97</td>
<td>3.98 ± 1.19</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>Oncocytoma</td>
<td>22</td>
<td>11.60 ± 14.41</td>
<td>3.42 ± 1.66</td>
<td>10.48%</td>
</tr>
<tr>
<td></td>
<td>Papillary RCC</td>
<td>21</td>
<td>10.49 ± 11.32</td>
<td>3.66 ± 1.47</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>Angiomyolipoma</td>
<td>12</td>
<td>7.98 ± 10.93</td>
<td>3.09 ± 1.44</td>
<td>5.71%</td>
</tr>
<tr>
<td></td>
<td>Chromophobe RCC</td>
<td>2</td>
<td>2.27 ± 1.15</td>
<td>2.20 ± 0.45</td>
<td>0.95%</td>
</tr>
<tr>
<td></td>
<td>Metastatic</td>
<td>2</td>
<td>18.95 ± 3.98</td>
<td>5.27 ± 1.12</td>
<td>0.95%</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>9</td>
<td>6.56 ± 11.37</td>
<td>2.69 ± 1.50</td>
<td>4.29%</td>
</tr>
<tr>
<td>Cyst</td>
<td></td>
<td>79</td>
<td>11.75 ± 32.03</td>
<td>2.67 ± 1.89</td>
<td>37.62%</td>
</tr>
<tr>
<td>All Masses</td>
<td></td>
<td>210</td>
<td>12.46 ± 3.34</td>
<td>3.34 ± 1.79</td>
<td>100%</td>
</tr>
</tbody>
</table>

Consideration of renal mass pathology is relevant to the kidney segmentation study because the same validation set is used for early stopping during hyperparameter tuning in both studies. The 118 NG-MRI volumes were divided into train, validation, and test sets based on an 80/10/10 split such that the validation and test sets contain 12 volumes each and the training set contains 94 volumes. This split format was chosen because a large training set was needed to train an
effective deep learning model. Recognizing that smaller test and validation sets could lead to biased results, steps were taken to ensure that these sets are representative of the distribution of renal masses in the entire dataset. Validation and test sets were randomly selected until a group of sets was found for which the renal mass distributions resemble the pathology distributions in Table 3. Chromophobe RCC and metastatic tumours were not considered for the validation and test sets since there were not enough cases to assign one to train, validation, and test sets. Solid masses with unknown pathology were also not considered for validation and test sets. These outlier cases were included in the training set as they still represent renal lesions. The pathological makeup of renal masses in the validation and test sets can be found in Table 4 and Table 5 respectively. Training, validation, and test set splits were kept constant for all phases of model development including renal mass detection. It was later also decided to complete final model evaluation on the entire 118 patient dataset using 5-fold cross validation.

<table>
<thead>
<tr>
<th>Renal Mass Type</th>
<th>Renal Mass Subtype</th>
<th>Number of Masses</th>
<th>Proportion of All Masses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solid Renal Mass</td>
<td></td>
<td>12</td>
<td>57.1%</td>
</tr>
<tr>
<td>Clear Cell RCC</td>
<td></td>
<td>6</td>
<td>28.6%</td>
</tr>
<tr>
<td>Oncocytoma</td>
<td></td>
<td>2</td>
<td>9.5%</td>
</tr>
<tr>
<td>Papillary RCC</td>
<td></td>
<td>2</td>
<td>9.5%</td>
</tr>
<tr>
<td>Angiomyolipoma</td>
<td></td>
<td>2</td>
<td>9.5%</td>
</tr>
<tr>
<td>Cyst</td>
<td></td>
<td>9</td>
<td>42.9%</td>
</tr>
<tr>
<td>All Masses</td>
<td></td>
<td>21</td>
<td>100%</td>
</tr>
</tbody>
</table>
### Table 5: Breakdown of test set by renal mass pathology.

<table>
<thead>
<tr>
<th>Renal Mass Type</th>
<th>Renal Mass Subtype</th>
<th>Number of Masses</th>
<th>Proportion of All Masses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solid Renal Mass</td>
<td>13</td>
<td>65%</td>
<td></td>
</tr>
<tr>
<td>Clear Cell RCC</td>
<td>7</td>
<td>35%</td>
<td></td>
</tr>
<tr>
<td>Oncocytoma</td>
<td>2</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Papillary RCC</td>
<td>2</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Angiomyolipoma</td>
<td>2</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Cyst</td>
<td>7</td>
<td>35%</td>
<td></td>
</tr>
<tr>
<td>All Masses</td>
<td>20</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

#### 4.1.2 CECT Dataset

This study explored the effect of cross-modality transfer learning from CECT data on model performance. Similar to the NG-MRI data, the CECT dataset was provided by TOH though a data sharing agreement with the University of Guelph and approved for use by the research ethics boards of both institutions. A total of 326 volumes were collected retrospectively between January 1st, 2015 and December 31st, 2017. All volumes in the dataset are comprised of contrast enhanced images in the nephrographic phase. Segmentations of kidney and renal masses were performed by a radiology fellow with corroboration with surgical and pathological reports. Despite the overlap in healthcare institutions, it is unknown if any of the CECT and NG-MR images are of the same patient due to anonymization. Since this CECT dataset is only considered for transfer learning purposes, the entire dataset is used as a training set. Details regarding the properties of renal masses in the CECT dataset are shown in Table 6 and an example of the provided segmentation can be viewed in Figure 7.
Table 6: Breakdown of renal mass types within the CECT dataset.

<table>
<thead>
<tr>
<th>Renal Mass Type</th>
<th>Number of Masses</th>
<th>Mean Volume (cm³)</th>
<th>Mean Diameter (cm)</th>
<th>Proportion of All Masses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solid Renal Mass</td>
<td>479</td>
<td>10.95 ± 64.70</td>
<td>3.72 ± 1.42</td>
<td>59.9%</td>
</tr>
<tr>
<td>Cyst</td>
<td>320</td>
<td>81.82 ± 197.42</td>
<td>5.47 ± 3.32</td>
<td>40.1%</td>
</tr>
<tr>
<td>All Masses</td>
<td>799</td>
<td>49.12 ± 157.19</td>
<td>3.84 ± 3.39</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 7: Axial view of the kidney in a CECT image (left), and its corresponding radiologist-provided segmentation where yellow indicates healthy kidney tissue, red indicates a solid renal mass, and blue indicates fluid-filled cysts (right).

4.2 Methods

The deep learning algorithm for segmenting kidney boundaries is described in this section. A high-level overview of the development process for the proposed model is shown in Figure 8. The model is first developed without considering transfer learning and then compared to the model with pre-training on the CECT data. Final model evaluation is completed on the entire NG-MRI dataset using 5-fold cross validation.
Figure 8: Process flow for developing the kidney segmentation algorithm.

4.2.1 Preprocessing Techniques

With the exception of image intensity normalization, the CECT and NG-MRI datasets were preprocessed identically. The intensity values of CT images can be interpreted using the Hounsfield unit (HU) scale. CECT images from various manufacturers were converted from their proprietary pixel representations to manufacturer-independent HU representations using the “Rescale Slope” and “Rescale Intercept” values within the DICOM header of each volume. Following this, a window-level function was applied with a level of 30 and a width of 310, a common range used in the KiTS19 challenge because it corresponds to HU values of abdominal organs in CECT images [81]. Finally, the intensities of each volume were normalized to be between 0 and 1 for compatibility with the deep learning model.

The intensity values of MR data does not map to a manufacturer-independent representation like CT data, so no window-level operation was used. Instead, each volume was normalized to be between the range of 0 and 1, with the maximum intensity being set to 1 and the minimum intensity being set to 0. An alternative method of normalization was tested where each image was normalized to be between the 5th and 95th percentile intensities. The alternative approach is theorized to reduce the effect of noise by excluding outlier intensities. Although, since no benefit in model performance was observed experimentally with the alternate method, it was decided to retain all image intensities during normalization. The following operations apply to both the datasets.

After intensity normalization, adaptive histogram equalization was applied because it was experimentally found to increase the average intensity difference between renal masses and healthy kidney tissue in the training set. Next, the image sizes were standardized by zero-padding.
and resampling using bilinear interpolation. To simplify compatibility with the U-Net architecture, images were standardized to shapes that could be downsampled multiple times using max-pooling operations. In the case of the 2D U-Net, axial images slices were resized to (256, 256) because of its similarity to the median axial image size, (224, 320) and its compatibility with memory restrictions of GPU units used in training. Zero-padding was used before resampling to ensure aspect ratios were preserved upon resizing. In the case of the 3D U-Net, similar padding and resizing operations were performed except the target volume size was set to (160, 160, 160). This size was chosen for training with the 3D volumes to ensure models can be trained within the 32GB memory limit of a V100 GPU.

The final step for preprocessing input images before training was data augmentation. Images were augmented by applying random elastic deformations to introduce more variation to the training set and reduce overfitting. Elastic deformation is a widely recommended augmentation technique for medical segmentation tasks because it represents morphological changes that are biologically explainable [36]. Other common augmentations used in medical image segmentation tasks include scaling, rotations, brightness and contrast changes [51]. This study only implemented elastic deformations since it is the most widely used augmentation in medical image segmentation and is unique because it morphologically deforms organ appearance in a physiologically explainable way. Further improvements to generalization may be made by including more augmentation techniques, but their effect on final image appearance was considered a risk. For example, augmenting by cropping or rotating was found to eliminate part of the kidney region on some slices where the organ is close to the image boundary.

The ground truth segmentations were preprocessed by converting them to binary segmentations where the background was set to 0 and the foreground set to 1. For kidney segmentation, the goal was to segment all components of the organ including the healthy kidney tissue, cysts, and solid masses. For this purpose, all non-background labels were set to 1. The segmentations were then resized to match their corresponding image, however nearest-neighbor interpolation was used instead of bilinear interpolation to preserve the binary nature of the segmentation.

4.2.2 Model Architecture

The deep learning models used to perform kidney segmentation were based off the U-Net architecture proposed by Ronneberger et al. [36]. There are many variations of U-Net in
literature today that have seen success in the task of kidney segmentation in CT through the KiTS19 challenge, although it has been established that optimizing model parameter configuration is often more rewarding than architectural variations [51][52]. A manual implementation of the U-Net was chosen over nnU-Net style semi-automated frameworks to emphasize interpretability and customizability in the development process. The two architectures considered in this study were the standard U-Net and the Attention U-Net, a model proposed by Schlemper et al. that incorporates attention gates at each up-sampling layer [45]. The Attention U-Net was chosen due to its theorized ability to suppress false positive detections. Both 2D and 3D configurations of these models were explored since strong results have been reported using models of both dimensionalities for kidney segmentation [61][65]. Additionally, the impact of model depth on segmentation performance was considered by testing different model sizes. Since the area of kidney segmentation in MRI is not well explored, the focus of my work was on establishing baseline performance with a well configured U-Net model. As a result, architectural testing was limited to the use of attention units, the dimensionality of the model, and the model depth. In each test, the best performing architecture on the validation set was selected to proceed. It should be noted that all architectural changes considered in this study share the common property of changing the number of tunable model parameters by modifying model hyperparameters. This makes the effectiveness of each change dependant on the number of training samples available, with larger models often requiring more training examples to converge. As a result, studies that use larger datasets may come to a different conclusion as to which architecture is the most suitable.

Diagrams of the four major model architectures can be seen in Figure 9 to Figure 12. U-Net models are arranged in a symmetric “U” shape, with the left branch serving as an encoder pathway that extracts increasingly higher level features. The right branch serves as a decoder which incorporates features from previous layers to generate a high resolution segmentation map. Intermediate convolutional layers have a kernel size of 3 and a stride of 1 in each dimension. Outputs from the convolutional layers are randomly set to 0 with a chance of 25% through a dropout layer to prevent overfitting. The dropout rate of 25% is a hyperparameter that was chosen experimentally. After the dropout layer, a rectified linear unit (ReLU) activation is applied which sets all negative outputs to 0. At the end of each encoder layer, a max-pooling operation with a stride length of 2 in each dimension reduces the size of all feature maps by half.
Conversely, at the beginning of each decoder layer, an up-convolution layer with a stride length of 2 in each dimension doubles the size of all feature maps. Each set of up-convolved feature maps is concatenated with feature maps of the same size from the encoder branch of model through skip connections. The parameters for kernel size and feature map size are modified versions of the approach taken by Zabihollahy et al. in [61], a recent study that attempted kidney segmentation and renal mass detection in CECT images. In the case of the Attention U-Net, feature maps from the skip connections are multiplied by a probability map generated from the upconvolved layer. The probability map contains values between 0 and 1 indicating the saliency of each feature. A final convolution layer at the end of the model converts the feature maps into a segmentation map. This layer has a kernel size of 1 in each dimension and uses a sigmoid activation function to ensure outputs are scaled to be between 0 and 1, allowing for computation of the loss by comparing to the ground truth.

Figure 9: Architecture of a 2D U-Net model with 6 layers of skip connections.
Figure 10: Architecture of a 3D U-Net model with 5 layers of skip connections.

Figure 11: Architecture of a 2D Attention U-Net model with 6 layers of skip connections.
Two choices for loss functions were considered for the kidney segmentation model, Dice loss and binary cross entropy (BCE) loss introduced in Eq. (5) and (4) respectively. Dice loss is often used in deep learning models performing kidney segmentation since DSC is the most widespread performance evaluation metric [61]. The BCE loss function incorporates logarithmic components that penalize predictions that are confident and wrong disproportionately higher than those that are only wrong by a small margin. This disproportionate penalization has a regularization effect that may result in a model that is more generalized. In the top submission to KiTS19, the authors used a summation of Dice and BCE (Dice-BCE) loss for optimal performance [75]. Dice-BCE loss may produce optimal results by combining the strengths of its components: the direct relationship Dice has with DSC-based evaluation and the regularizing effect of BCE loss.

Dice loss, BCE loss, and Dice-BCE loss are tested using 5-fold cross validation to determine the best loss function for the kidney segmentation in NG-MRI task. Cross validation testing was used for determining the best loss function for kidney segmentation because these changes are expected to result in subtle improvements to performance that may be more vulnerable to biases in the validation set. While it would be ideal to use cross validation for all model selection tests, its use was limited to where it may have the highest impact in order to minimize resource usage.
4.2.3 Transfer Learning

The CECT dataset is significantly larger than the NG-MRI dataset and contains potentially unique kidneys and renal masses that are segmented using the same strategy. I attempted a transfer learning based approach to improve kidney segmentation performance in NG-MRI by first pre-training the optimal kidney segmentation model to complete an identical task in CECT images. After training on the CECT dataset, the feature extractors of the model may converge to states that are favorable to the kidney segmentation task in NG-MRI. This convergence is expected because signals from both modalities are influenced by the morphological shape of the kidney. Additionally, both images are contrast enhanced in the nephrographic phase which will make the renal medulla region hyperintense, further increasing morphological correlation between the images. This form of transfer learning, referred to as full network adaption, requires the model architecture to be kept identical between tasks so that every parameter can be pre-trained [57]. To determine which model would be tested with transfer learning, the model that achieved optimal performance on the NG-MRI dataset with randomly initialized weights was selected. This was done to ensure architecture selection decisions are made to optimize performance on the NG-MRI dataset, not the CECT dataset. Another common approach for transfer learning involves freezing the weights of certain layers after pre-training and only allowing certain parts of the model to be fine-tuned [57]. Such an approach is ideal when there is a large degree of similarity between the two tasks and initial layers of the model are effective in both. Since CECT and NG-MRI are fundamentally different modalities that are subject to different types of noise and artifacts, partial adaption was not considered.

The benefit of transfer learning is known to depend on the size of the training data available in the target task [57]. If there is ample training data available in the target task, minimal improvement will be seen through transfer learning compared to random weight initialization. When data in the target task is limited however, pre-trained weights can improve model performance. In this study, the target task contains 94 NG-MRI volumes in the training set. Since it is not known if this is a large or small amount of data in the context of transfer learning, the benefit of full network adaption is measured when the dataset size of the target task is artificially reduced. This will help establish a threshold of ground truth segmentations in a target task of a different modality after which cross-domain transfer learning benefits diminish.
4.2.4 Model Development and Training

The deep learning models developed in this thesis were implemented in Python using the TensorFlow library with a Keras functional-API [82]. Pre-processed medical images and ground truth segmentations were saved as Numpy files and loaded by TensorFlow at run-time using a data generator. This approach had several benefits. Instead of loading all images into system memory simultaneously, the data generator only loaded enough images to fill one batch for training. While the loaded batch was being processed by the machine learning model, the data generator used a prefetch function to load the next batch of images which reduced over all training time. Images within a batch were selected by a random shuffling process to prevent the order of batches from being identical every epoch. Furthermore, the data generator applied data augmentation to input images at run-time through a random process. Each loaded image in a batch had a 25% chance of being augmented, otherwise it was shown to the model in its original form. This approach reduced storage requirements by eliminating the need to store augmented versions of images and increased variation in the input images by applying a random augmentation every time.

Training for this study was completed on Nvidia P100 and V100 GPUs located on the Graham and Cedar clusters of the Compute Canada network. Using a high performance computing cluster benefited the training process by allowing tasks to be parallelized. For example, to implement 5-fold cross validation 5 different GPUs were requested simultaneously, effectively reducing cross validation time by 80%. The 2D kidney segmentation models were trained in batch sizes of 36 and 3D kidney segmentation models were trained in batch sizes of 2. These sizes were determined based on model sizes and memory restrictions of the GPU. After each epoch of training, the validation loss was calculated, and the model weights were stored. If the validation loss did not improve after 10 epochs, training was stopped and the set of weights that performed the best on the validation set were stored. This process of early stopping allows the number of epochs of training to be dynamically set based on loss convergence and reduces the risk of overfitting the training set. All loss functions used the Adam optimizer with a learning rate of $1 \times 10^{-3}$ to tune weights during training. This learning rate was found during hyperparameter tuning to result in stable loss convergence while minimizing training time. The Adam optimizer
further tunes the learning rate on a per-parameter basis during training based on the magnitudes of previous parameter updates [47].

4.2.5 Postprocessing Techniques

The output kidney segmentations from the deep learning model contained values ranging from 0 to 1 due to the sigmoid activation layer. These outputs were converted to a binary segmentation map by applying a threshold at 0.5, a value taken from previous studies that also worked well on the NG-MRI dataset [61]. Intensity thresholding was implemented by setting any intensity at or above 0.5 to 1 and any intensity below 0.5 to 0. After thresholding, 2D segmentations were compiled to 3D volumes, where blob detection was used to improve results. Since it is known that the kidney segmentation is a connected component in 3D space, an effective kidney segmentation algorithm will represent the kidneys as the largest connected components in a binary 3D segmentation mask. This information is leveraged in postprocessing by eliminating any components of the segmentation that are not spatially connected to the two largest spatially separated blobs as follows.

A Laplacian of Gaussian (LoG) blob extractor was used to extract 3D blob locations in the predicted segmentation mask. The LoG blob extractor provides not only the location of a blob, but the size of the Gaussian kernel used to detect a blob which correlates with the size of a detected blob. Hence, the detected blobs were ordered from largest to smallest based on the kernel size they were detected at and the two largest were chosen as locations for the kidney. These locations were then used as seed points for a connected component algorithm that defined the pixel-wise boundary of each kidney blob, and any region in the binary mask that is not part of the two connected components was eliminated. While stand-alone connected component algorithms exist that do not use LoG to define seed points, the LoG extractor was used because it was an intuitive means of choosing seed points that were observed to work well on binary masks. Note that this approach would work for a patient that only has one kidney, although in that scenario the kidney segmentation is expected to be more prone to false positives. To better deal with these cases in a practical implementation, detected blobs may be filtered by a threshold for reasonable kidney volume.
4.2.6 Evaluation of the Models for Segmentation Accuracy

Segmentation accuracy was measured in 3D against the validation set of the NG-MRI data during training. The training phase determined optimal model architecture, loss function configuration, and transfer learning use. After training, final model evaluation was completed on a test set composed of 12 NG-MRI volumes of unseen data, and on all available data through 5-fold cross validation. All evaluation metrics for segmentation were computed on a patient-wise basis and averaged for all patients in a set. DSC score was the primary method of assessment in the training phase, because it is the most common means of assessing segmentation accuracy in literature [51]. In addition to DSC, voxel-wise precision and recall data was recorded directly to observe trends not captured by the DSC score. Accuracy of segmentation was also measured based on the Hausdorff Distance (HD) as per Eq. (8). The HD metric finds the point on the predicted boundary in 3D that is the farthest away from any point on the ground truth and averages this distance across all cases tested. It provides a measure of the greatest segmentation error seen on average. A low HD score would indicate the kidney segmentation model is consistent and accurate at the boundary level in all anatomical regions of the kidney. Finally, kidney segmentation performance is also measured based on relative absolute volume difference (RAVD) as per Eq. (9). In contrast to HD, RAVD measures the aggregate performance of the segmentation. Minor errors in over or under segmentation of the boundary would not have a strong impact on RAVD but missing an entire region or area of significant volume would.

These evaluation metrics together provide an overview of kidney segmentation model performance, however in the context of this thesis the main purpose of kidney segmentation is to localize the kidney region for the renal mass detection algorithm proposed in the next chapter. Under-segmentation errors in the predicted kidney segmentation may remove regions corresponding to renal masses and result in a lower detection recall score. Conversely, over-segmentation may include tissue outside the kidney that appears similar to a renal mass and result in a lower detection precision. A final assessment of the kidney segmentation performance will therefore be made in Chapter 5, when the renal mass detection performance using ground truth kidney segmentations is compared to using the model generated kidney segmentations.
4.3 Results

The first set of results compared the accuracy of four different architectures with a depth of 5 layers, a Dice loss function, and hyperparameters configured as described in section 4.2.2. These parameters were chosen as starting points based on the work of Zabihollahy et al. in [61]. The performance of each architecture on the validation set after training with early stopping can be found in Figure 13. All models performed similarly, however the 3D models generally produced worse results based on every metric except for Hausdorff distance. The 3D models also took considerably more resources to converge than their 2D counterparts. A standard 3D U-Net converged in 592 epochs resulting in a total training time of approximately 63 hours using a V100 GPU. In contrast, a 2D standard U-Net converged in 38 epochs with a training time of approximately 9 hours using a P100 GPU. This difference in convergence time can be attributed to the sizes of the models. The 5-layer standard 2D U-Net had 7,775,313 parameters whereas the 5-layer standard 3D U-Net had 22,620,529 parameters. Based on its inferior performance and long training time, 3D U-Net models were not considered for the kidney segmentation task.

Notably, the Attention U-Net did not significantly outperform the standard U-Net, with the 2D standard U-Net reporting a DSC of 89.99% ± 7.60% and the 2D Attention U-Net reporting a DSC of 89.77% ± 4.99%. It was decided to proceed with a standard 2D U-Net for further testing since it produced similar results to the Attention U-Net while having a less complex architecture.

The optimal depth of the 2D U-Net for kidney segmentation was determined to be 6 layers through testing seen in Table 7. A 6 layer model reported the highest DSC score, which is the most significant performance metric for this task. Additionally, it reported the lowest RAVD indicating that the kidney volume was captured accurately. The 7 layer model reported the lowest HD value, however its inferior result in the other two metrics lead to decision to exclude this model from further testing.
After determining the ideal model architecture and depth, multiple loss functions were tested using 5-fold cross-validation as per Table 8. The combination of Dice and BCE loss resulted in the highest DSC score, making Dice-BCE loss the optimal choice for the 2D U-Net model for kidney segmentation. Notably the BCE loss model reported slightly better performance than
Dice-BCE in terms of RAVD and HD. Despite this, the Dice loss component was included in the final model to retain relevance and achieve better performance in terms of DSC.

Table 8: Kidney segmentation accuracy of 2D U-Net models with varying loss functions. Results are reported on a superset of the training and validation sets through 5-fold cross validation.

<table>
<thead>
<tr>
<th>Loss function of the model</th>
<th>DSC (%)</th>
<th>RAVD (%)</th>
<th>HD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dice loss</td>
<td>88.48 ± 9.70</td>
<td>10.87 ± 11.81</td>
<td>26.26 ± 29.69</td>
</tr>
<tr>
<td>Dice-BCE loss</td>
<td><strong>91.02 ± 5.59</strong></td>
<td>7.06 ± 8.06</td>
<td>21.34 ± 27.86</td>
</tr>
<tr>
<td>BCE loss</td>
<td>89.12 ± 4.77</td>
<td><strong>6.78 ± 8.83</strong></td>
<td><strong>19.18 ± 23.11</strong></td>
</tr>
</tbody>
</table>

The performance of this model on the CECT dataset was then considered in the context of transfer learning. Comprehensive analysis of similar deep learning models performing kidney segmentation in this CECT dataset was already completed by Zabihollahy et al. in [61]. An attempt was made to verify that the kidney segmentation model designed in this thesis achieves comparable accuracy to [61] by training a kidney segmentation model on 90% of the CECT dataset and using the remaining 10% as a test set. The results of this verification seen in Table 9 show that the U-Net model designed in this study is highly effective at kidney segmentation in the CECT dataset and achieves similar DSC performance to [61]. Further experiments with transfer learning were completed by training the model on all 326 CECT cases. To test the effect of transfer learning from CECT, the optimal kidney segmentation as determined by the previous tests was first trained on the CECT dataset and then fine-tuned on the NG-MRI dataset.

It was found that transfer learning did not result in a significant improvement in DSC score on the NG-MRI validation set compared to a randomly initialized model when the entire NG-MRI training set was used for fine-tuning. As a result, the final NG-MRI kidney segmentation model did not incorporate weights pre-trained on the CECT dataset. Notably, transfer learning did result in an improvement in DSC score when the number of cases in the NG-MRI dataset was artificially restricted as seen in Figure 14. With just 7 NG-MRI training cases, the transfer learning based kidney segmentation model reported a DSC score of 87.87% ± 12.77% on the validation set. This result indicates that cross-modality transfer learning may be used to fine-tune a model to perform kidney segmentation in a different domain with minimal training data.
Table 9: CECT kidney segmentation accuracy of the U-Net model compared to the U-Net model proposed by Zabihollahy et al. in [61].

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Set Size</th>
<th>DSC (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>RAVD (%)</th>
<th>HD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D U-Net 32 CECT volumes</td>
<td>94.50 ± 4.14</td>
<td>60.01 ± 3.34</td>
<td>94.18 ± 6.61</td>
<td>4.68 ± 6.82</td>
<td>9.09 ± 8.91</td>
<td></td>
</tr>
<tr>
<td>2D U-Net from [61] 125 CECT volumes</td>
<td>95.79 ± 5.16</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Figure 14: Effect of transfer learning from CECT on NG-MRI kidney segmentation DSC based on the NG-MRI training set size.

The final model for kidney segmentation in NG-MRI was a 2D U-Net with a depth of 6 layers and a Dice-BCE loss function. The architecture of this model can be seen in Figure 9. Test set results were acquired by training the deep learning model on a superset of the training and validation sets and evaluating performance on the test set. Based on the results seen in Table 10,
the final NG-MRI kidney segmentation model is highly effective with a DSC of 94.48% ± 1.11% on the test set. This model converged in 58 epochs resulting in a total training time of 16 hours. Fully autonomous kidney segmentations were produced for the test set with an average computation time of 2.13 minutes per patient. The axial segmentation accuracy can be visualized in 2D slices in Figure 15. From these images it can be observed that kidney segmentation accuracy is dependant on the anatomical region, with the kidney base being segmented well. The apex regions of the kidney show relatively poor segmentation accuracy, attributable to the ambiguous appearance of the organ boundary in these slices. The regions that generate the largest HD error often corresponded to the renal hilum, an area where the vasculature and ureter leave the kidney. This area may be challenging to segment because its appearance is significantly different from the renal medulla. Notably, the KiTS19 challenge excluded the renal hilum in their ground truth, making it a less complex segmentation task [51].

Based on the 3D renderings in Figure 16, it can be observed that the 3D shape of the organ is effectively captured by the deep learning algorithm despite segmentations being generated on 2D slices. Moreover, the 3D surface of the predicted segmentation appears smoother than the ground truth segmentation. The rough edges of the ground truth segmentation may be attributable to minor errors in the segmentation labeling process by the clinical expert. Some differences in the 3D surface texture may also be the result of over-simplification of the kidney boundary by the algorithm. Observing the bottom left predicted segmentation image in Figure 16 and comparing it to its corresponding ground truth segmentation, it can also be observed that the bottom apex region is under-segmented. This type of under segmentation error may result in reduced renal mass detection recall if a renal mass were to be located there. Despite these errors, observing the ROC plot shown in Figure 17 and the associated AUC value of 0.972 indicate that the final kidney segmentation model is highly accurate.

Table 10: Kidney segmentation performance of the final U-Net model. Results are reported on the NG-MRI test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>DSC (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>RAVD (%)</th>
<th>HD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D U-Net</td>
<td>94.48 ± 1.11</td>
<td>94.02 ± 1.91</td>
<td>94.99 ± 1.72</td>
<td>2.67 ± 1.96</td>
<td>8.58 ± 2.86</td>
</tr>
<tr>
<td>2D U-Net 5-Fold Results</td>
<td>91.20 ± 5.41</td>
<td>90.56 ± 5.95</td>
<td>92.67 ± 10.65</td>
<td>7.89 ± 6.08</td>
<td>20.67 ± 26.62</td>
</tr>
</tbody>
</table>
Figure 15: NG-MRI Kidney segmentation prediction axial views. The ground truth segmentation is in green and predicted segmentation is in red. Each column contains images from a different patient in the test set. The first row corresponds to a typical kidney base segmentation result, the middle row corresponds to a typical kidney apex segmentation result, and the last row corresponds to the slice that generated the maximum Hausdorff distance error.
Figure 16: 3D views of the NG-MRI kidney segmentations. Each column corresponds to a different patient from the test set. Segmentations generated by the deep learning model in the bottom row (red) are compared to the ground truth segmentations in the top row (green).
4.4 Discussion

In this chapter, a fully automated kidney segmentation algorithm in NG-MRI has been described. Based on a DSC score of 94.48% ± 1.11% on the test set and a DSC score of 91.20% ± 5.41% on the entire NG-MRI dataset, the proposed kidney segmentation algorithm is highly accurate at delineating the kidney boundaries. Furthermore, considering an average RAVD of 7.89 ± 6.08, the model generally effectively captures the kidney volume. The average HD was approximately 2cm on the entire NG-MRI dataset, which can be interpreted as an average lower bound for segmentation performance in challenging anatomical regions such as the renal hilum and apex.

This work represents a significant contribution to the field of deep learning based medical image segmentation, where work on kidney segmentation in MR images is limited and no studies have attempted deep learning based kidney segmentation in contrast-enhanced MRI in adult patients. Considering that there is increased interest in using MR imaging for renal mass assessment [5], this work provides valuable insights on the effectiveness of deep learning algorithms in isolating the kidney, a key first step in automatically detecting renal masses. Compared to previous studies, the strength of my work lies in the explicit inclusion of tumour containing organs and the diversity of a multi-institution dataset with images from multiple MRI manufacturers in the contrast enhanced NG phase. The work of Langner et al. in [40] reported a DSC of 95.6% on
kidney segmentation in non-contrast enhanced T₁ weighted MR images from a single Siemens Aera 1.5T machine. The presence of renal masses in their dataset was not established, making comparison with the results of this study difficult. Similarly, Isensee et al. in [70] reported a DSC score of 94.6% in the CHAOS segmentation challenge, but their study excluded all organs that contained renal masses. The kidney segmentation results in this thesis therefore provide a novel perspective on deep-learning based kidney segmentation in renal-mass containing kidneys in MR images.

Furthermore, the developed algorithm is fully automatic and able to identify kidney boundaries from all axial slices of a clinical abdominal MR volume of varying size. These axial slices often spanned the pelvic and respiratory system region of the body yet did not produce any false positives that were disconnected from the main kidney mass components. This presents a significant advantage over the work of Shehata et al. in [68] which used shape-priors with a level-set based method and required images to be manually localized to the kidney region. The ability to autonomously isolate the kidney boundary while accommodating natural variations present in abdominal MR images is critical to the final goal of fully autonomous renal mass detection and will play a role in reducing the complexity of the detection task.

The results of transfer learning from CECT provide a secondary contribution to the field by reporting the effect of cross-modality pre-training for kidney segmentation, which has not been attempted before. Pre-training on the CECT dataset was not included in the final model since there was no improvement over a randomly initialized model. However, when the size of the NG-MRI dataset was artificially reduced, a noticeable improvement was observed. This result indicates that models trained on identical tasks in CECT could be adapted to their counterparts in contrast-enhanced MR with fewer expertly annotated images. These results follow the same trend seen in cross-modality studies in other areas such as the work of Kang et al. in [58] for liver segmentation, where fine-tuning with as few as 10 images in the target modality produced an accurate segmentation. Acquiring annotated images is a significant bottleneck in medical image segmentation studies which could be partially mitigated by a transfer learning approach. This is particularly relevant to studies on the kidney where large CECT datasets and small MRI datasets are the norm.
The deep learning algorithm described in this chapter was based on the U-Net architecture originally proposed by Ronneberger et al. in [36] and has seen widespread adoption in similar kidney segmentation tasks [51]. Testing revealed that the standard U-Net architecture matched the performance of the Attention U-Net while having a simpler architecture. This agrees with the general sentiment that a simple U-Net model that is well configured is more relevant to strong segmentation results than architectural changes, and that configurations are highly domain and task dependent [52]. Additionally, the testing revealed that a 2D U-Net outperformed a 3D U-Net, a deviation from the leading results in KiTS19 and some other studies on deep learning based kidney segmentation [40, 52, 33]. All of these studies had larger datasets than this study, which may explain their preference for a 3D U-Net since it has considerably more parameters than its 2D counterpart, and as such requires more data and resources to train. Competitive results have been published recently using 2D U-Net as well, as seen in [61], indicating that the choice of 2D or 3D architectures may be domain, task, and dataset specific.

A limitation of this kidney segmentation study was that ground truth segmentations were only provided by one clinical expert. This ground truth directly influenced model development and evaluation results. Any biases or variability in the organ segmentation by the clinical expert may therefore be reflected in the results of the study. Other studies such as KiTS19 have commonly used segmentations from multiple experts which can be used to establish a measure of inter-operator variability [51]. Furthermore, the clinical expert in this study had access to multi-parametric imaging data for each patient when producing segmentations, while the deep learning model was tested using only the NG-MRI data. As a result, the ground truth segmentation may be influenced by information not apparent in the NG-MRI volume, making the evaluation metrics an inherently conservative estimation of the kidney segmentation performance of the deep learning model. A model that can incorporate information from all sequences available for a patient may achieve better segmentation accuracy and could be compared more fairly to the ground truth data, representing a potential future direction for improving the kidney segmentation results. A final limitation to consider is the size of the NG-MRI dataset itself. Compared to similar studies in CECT, the small dataset may not be as strong a representation of the variability seen in real world data. While efforts were made in this study to maximize training dataset size while making the evaluation a fair representation of the available data in
terms of pathology, further studies with larger NG-MRI datasets may make a better assessment of the real world performance of deep learning based kidney segmentation algorithms.

4.5 Conclusion

The described algorithm was found to be effective at autonomously producing kidney segmentations in NG-MRI volumes. The final algorithm is a 2D U-Net with a depth of 6 layers and a Dice-BCE loss function. Attempting transfer learning from CECT revealed that cross-modality pre-training can be valuable in situations where data in the target domain is limited. In this study pre-training on CECT lead to a significant improvement in performance when the target training set size was limited but had negligible benefits when the entire NG-MRI training set was used. The proposed algorithm is capable of serving the purpose of a localizer for the main objective of this thesis, fully autonomous renal mass detection.
Chapter 5: Development of an Algorithm for Renal Mass Detection

The objective of this chapter is to propose and test a fully autonomous algorithm for detecting renal masses within NG-MRI images. Renal masses include fluid-filled cysts and solid masses containing benign or malignant tumours. Computer aided detection of renal masses in MRI has not been attempted in literature despite increased interest by the clinical community in using MRI for renal mass assessment [14]. Multiparametric MRI can improve outcomes for patients by providing a more accurate assessment of malignancy than CECT while being non-invasive and non-ionizing [4]. This study contributes to the field of computer aided detection by developing a fully autonomous renal mass detection algorithm which has the potential to increase the detection sensitivity and specificity of diagnostic imaging using MRI. Furthermore, an effective renal mass detection algorithm can serve as a critical first step for a renal mass classification algorithm which can determine the pathological subtype of detected masses and provide additional clinical value. The work in this chapter is supported by the kidney segmentation algorithm developed in Chapter 4 which is used to reduce the complexity of the detection task by isolating the kidney sub-region.

5.1 Methods

An overview of the final algorithm for detecting renal masses is shown in Figure 19. Localized images of the kidney are acquired using the boundary generated by the kidney segmentation algorithm. These images are then fed into two independent algorithms, one for renal mass segmentation and another for healthy tissue segmentation. The outputs of these two models are ensembled using mean voting and a binary threshold is applied. Finally, the ensembled output is compiled into 3D volumes for each patient and spatially separated 3D masses are identified. The development process for the renal mass detection algorithm is shown in Figure 18. Hyperparameter tuning and transfer learning testing is only completed for the renal mass segmentation algorithm. At the ensemble stage, RM detection is evaluated based on fully autonomous data and on the entire NG-MRI dataset using 5-fold cross validation.
5.1.1 Preprocessing Techniques

Images for this study were obtained from the same NG-MRI and CECT datasets described in Section 4.1. Training, validation, and test set splits were kept identical to Chapter 4 to ensure renal mass subtypes are proportionally represented. These splits were used to find optimal hyperparameters for the renal mass and healthy tissue segmentation models, however, all evaluation related to fully autonomous renal mass detection was completed on the entire NG-MRI dataset using 5-fold cross validation. The first step in pre-processing images was localization to the kidney region. This localization was accomplished by using the model-generated and ground truth kidney segmentations. The kidney segmentation model reported a DSC of 91.20% ± 5.41% on the NG-MRI dataset, indicating that it generates an accurate but noisy estimation of the true kidney sub-region. Minor errors in localization may interfere with
the training process of the renal mass and healthy tissue segmentation algorithms, so training and validation images were localized using the ground truth kidney segmentations instead. In contrast, images that were used to assess fully autonomous detection accuracy were localized using the model-generated kidney segmentations.

Using a previously generated segmentation to localize a sub-region for a specialized segmentation task in a cascaded model approach is a common technique in previous studies [83][29]. In terms of renal mass segmentation, cascaded U-Nets have been applied but there is no consensus on the best localization method. The work of Zabihollahy et al. in [61] and Lin et al. in [65] used model-generated segmentations to extract bounding boxes which contained the kidney and surrounding tissue. Hou et al. in [76], which achieved the second-best result in KiTS19, used kidney segmentations to isolate the kidney subregion and eliminated all intensities not within the bounds of the kidney segmentation, leaving no background tissue in the image. Another submission to KiTS19 by Zhang et al., which achieved fourth place, used kidney segmentations to isolate the kidney subregion while keeping background intensities, but provided the model-generated kidney segmentations to the renal mass detection algorithm in a separate channel as a ‘spatial-prior’ [51]. This style of localization was also used more recently for prostate tumour segmentation in [30]. Only the localization techniques used by Hou et al. in [76] and Zhang et al. in [51] were considered in this study since they leverage rich semantic information from the kidney segmentation to provide context. These localization techniques will be referred to as background-eliminated and spatial-prior methods respectively in this thesis. Examples of these two localization methods are shown in Figure 20. Prior to testing different model architectures, both localization methods are attempted on the healthy tissue and renal mass segmentation tasks using a standard 2D U-Net. The method that scored the highest detection accuracy on the validation set is selected. Both localization techniques are also compared in the context of fully autonomous renal mass detection to determine how each localization method responds to under-segmentation errors in the kidney boundary.

Pure bounding box techniques that do not provide semantic level information regarding the kidney boundary were not considered because they do not provide the same level of context to a renal mass detection algorithm. Without this context, renal mass detection algorithms are expected to be more susceptible to false positive detections in the surrounding tissue. Notably,
none of the top five submissions to KiTS19 described in [51] used a pure bounding box style localization method for binary renal mass segmentation.

![Background-eliminated localization example](image1)

![Spatial-prior localization example](image2)

**Figure 20:** a) Background-eliminated localization example with an image and its ground truth, b) Spatial-prior localization example with a 2-channel input of the image and kidney segmentation and its corresponding ground truth.

The first step in preparing the original DICOM volumes for renal mass detection was to split each volume along the sagittal plane to effectively separate the left and right kidneys of a patient. The kidney segmentation was then used to define a 3D bounding box around the kidney and the rest of the image is cropped. Background-eliminated images were produced by setting all intensities outside the kidney segmentation boundary to 0. Backgrounds were retained for the spatial-prior method. In both cases, the images were normalized to be between 0 and 1 and adaptive histogram equalization was applied to increase the intensity contrast between tumours and healthy kidney tissue. Finally, axial images were rescaled to (128, 128) with a series of zero-padding and bilinear interpolation operations identical to the preprocessing steps for kidney segmentation. This image size was chosen because it closely resembles the median image size
after localization, (90, 72), and was compatible with the 2D U-Net. For compatibility with the developed 3D U-Net model, localized volumes were resized to (160, 160, 160). Data augmentation for images used in the renal mass detection study followed the same strategy described in Section 4.2.1. Elastic deformation was randomly applied to the training data at run time to simulate the effects of local tissue compression and expansion due to patient movement and breathing.

The ground truth renal mass segmentations were obtained by converting the NiFTI segmentation files into binary segmentations. In the case of renal mass segmentation, all labels belonging to renal masses were set to 1 and labels corresponding to healthy kidney tissue and background tissue were set to 0. For healthy tissue segmentation, labels corresponding to the background and any renal mass were set to 0, and the remaining kidney tissue was set to 1. The binary segmentation masks were then split along the sagittal planes and resized identically to their corresponding images to prepare them for the deep learning model.

5.1.2 Model Architecture for Renal Mass Detection

While the objective of this chapter is to detect renal masses, the deep learning algorithms are trained to perform segmentation tasks. This approach is common in studies that have attempted fully autonomous renal mass detection [61][65]. To further understand why a semantic segmentation model is the architecture of choice for a detection task, which has specialized architectures of its own, unique aspects of medical imaging datasets must be considered. It has been proposed by Jaeger et al. in [84] that models which incorporate segmentation tasks excel at being data-efficient object detectors because they are able to fully exploit the semantic segmentation signal. The goal of deep learning techniques in the medical imaging domain is often to perform data-efficient training due to their relatively small datasets such as the 118 image set used in this study. If object detection based architectures, such as the YOLO architecture [85], were to be used on the NG-MRI dataset, the fine-grain renal mass segmentations must be converted to rectangular bounding boxes. These bounding boxes would result in a loss of information since they would inevitably contain intensities from non-renal mass tissues, resulting in weaker foreground-background separation. The data-efficiency of semantic segmentation models is not as relevant to large non-medical object detection datasets such as ImageNet, which contains over 1 million images [34]. Therefore, the core deep learning
models for the renal mass detection task were specialized to perform segmentation but evaluated based on detection.

The segmentation models were developed in two phases, renal mass segmentation and healthy kidney tissue segmentation. Background-elimination and spatial-prior based localization techniques were considered independently for both the tasks. The focus in both phases was on maximizing the recall metric for RM detection. In the first phase, an optimal RM segmentation algorithm was developed with a focus was on determining what model architecture, depth, and loss function configuration results in the highest RM detection recall. Choices for the RM segmentation architecture were informed by testing completed in Chapter 4 for kidney segmentation, which found that a 6-layer 2D U-Net with a Dice-BCE loss function resulted in optimal results. This model was used as a starting point for the RM segmentation task because the kidney segmentation task also included renal masses so it is expected that many of the same architectural and hyperparameter choices will be relevant to RM segmentation. To accommodate for any differences in the segmentation tasks and verify that the chosen architecture is ideal for RM segmentation, the same series of tests performed for kidney segmentation are repeated here using the validation set. Attention U-Nets are tested in the 2D and 3D form as well as the standard U-Net in the 3D format, and model depth is verified by testing models with fewer and greater number of layers.

Finally, the optimal loss function is found using 5-fold cross validation on a combination of the training and validation sets. Loss functions considered for this task include Dice loss, BCE loss, and $F_\beta$-BCE loss. Dice-BCE loss was the best performing loss function on the kidney segmentation task. $F_\beta$-BCE loss extends this loss function by replacing the Dice component with the $F_\beta$ term described in Eq. (6). When an $F_\beta$ component is used within the loss function, the $\beta$ parameter must be determined as well. Based on the $F_\beta$ equation, choosing $\beta = 1$ causes the $F_\beta$ term to simplify to the Dice loss equation [49]. Furthermore, choosing a $\beta$ value less than 1 biases the loss function to penalize precision more, and choosing a $\beta$ value greater than 1 will bias towards penalizing recall more. Since the strategy of renal mass segmentation in the context of this study is to have a high recall, $\beta$ values at or above 1 were considered for renal mass segmentation and values below 1 were considered for healthy tissue segmentation.
The next phase of developing the ensemble model for renal mass detection involved finding an optimal healthy kidney tissue segmentation model. Since this model finds where there are no renal masses, optimizing it for high precision has the effect of increasing recall for renal masses. The healthy tissue segmentation model can be used independently to detect renal masses by converting its output to a renal mass segmentation. This conversion process entails subtracting the output of the healthy tissue model from the output of the kidney segmentation model. The remaining region in the kidney segmentation after subtraction corresponds to the predicted renal mass segmentation boundaries. The optimal model architecture from the renal mass segmentation task was used for the healthy tissue segmentation task. Based on the complimentary nature of both tasks, it is expected that most of the hyperparameters found for the renal mass segmentation task will be applicable here, with the exception of the localization technique and loss function optimization. Both of these aspects are tested for independently in the context of healthy tissue segmentation. When determining the ideal loss function for healthy kidney tissue segmentation, BCE loss, Dice loss, and $F_β$-BCE loss with $β$ values at or below 1 were considered using 5-fold cross validation.

The ensemble model for renal mass detection uses a mixture-of-experts (MoE) style architecture which contains the same number of components as there are labels in the output [62][86]. Since the output of the ensemble is binary, this essentially means two components are used. One component is a renal mass segmentation model, which performs binary segmentation to identify the locations of renal masses. Upon trial of independently using the renal mass segmentation for detection, it was found that the renal mass segmentation was overly sensitive resulting in poor recall. It was then decided to train a separate model to perform healthy tissue segmentation. The healthy tissue segmentation model also performs binary segmentation, although its goal is to label healthy tissue as the foreground. These two tasks, renal mass segmentation and healthy kidney tissue segmentation, are complementary to each other but are non-intersecting, making them ideal for an MoE approach. This is in contrast to the ensemble approach used by Zabihollahy et al. in [61] where multiple models are trained to perform the same task and predictions are combined using majority voting. Conventional MoE approaches often combine expert predictions using a trainable gating network [62], primarily because the interaction between experts can be complex. However, in this case the interaction between the outputs of the
experts is simple since they are compliments of each other. Two interpretable methods of combining outputs are therefore considered, mean averaging and healthy tissue gating.

The healthy tissue gating method of combining outputs was accomplished by using the healthy tissue segmentation as a screen that eliminates false positive regions in the renal mass segmentation. The renal mass segmentation alone was found to have high recall but many false positives. By training the healthy kidney tissue segmentation model to have high precision, I attempted using its output directly as a gating function to suppress false positive detections in the renal mass segmentation. The healthy tissue segmentation task is well suited for achieving high precision since healthy kidney tissue can be expected to show a high degree of uniformity, with the exception of the renal hilum which is a diverse region containing the ureter and vasculature. In contrast, renal masses are known to contain many subtypes which are defined by their unique visual appearances [4]. This ensemble approach takes advantage of the observed over-sensitivity of the renal mass segmentation algorithm and the high precision of the healthy tissue segmentation to attempt to increase renal mass detection accuracy.

The mean averaging method of creating an ensemble uses pre-threshold outputs of each component. Healthy tissue segmentations are converted to renal mass segmentations by subtracting them by one and taking the absolute value. As a result, its output is converted to the same task as the renal mass segmentation. The two outputs were then simply averaged together by taking the pixel-wise mean. A binary threshold was then applied to the ensemble output at a value of 0.5, creating the final renal mass segmentation prediction.

5.1.3 Detection of Spatially Separated Instances

Once boundaries of renal masses in an image were identified, spatially separated instances of renal masses were found in 3D. To initialize this process, a Laplacian of Gaussian (LoG) blob detector was applied to the image. The LoG detector outputs centroids which are used as initialization points for a connected component algorithm. This detection method was initially chosen because setting the minimum size of the Gaussian kernel applied to the segmentation map enables the rejection of small blobs, which often correspond to false positives. In practice, this features usefulness is limited to very small speckle-like blobs since the developed models often segmented incomplete regions of renal masses which result in blobs that are significantly smaller than the true renal mass diameter. Despite this limitation, the LoG detector was found to be
effective at suggesting initialization points for the connected component algorithm and provided value by ordering blobs based on their size. Using connected components is beneficial because it uses the binary prediction output to define boundaries of each detected instance. This boundary is compared with the ground truth based on the spatial-overlap method described in Eq. (10). Any predicted mass that overlaps with a ground truth blob by at least one voxel is counted as a true positive. If a predicted blob does not overlap with a blob in the ground truth, it is counted as a false positive. Conversely, a ground truth mass that does not overlap with any predicted mass was counted as a false negative. If multiple predicted masses intersected with the same ground truth mass, they were still only counted as one true positive and the multiple detections were not counted as false positives. This strategy ensures the model is not unfairly rewarded for detecting a single renal mass as multiple disconnected masses. The impact of detecting a mass in multiple locations in a real world scenario is expected to be limited since each detection must have some overlap with the true renal mass location. Furthermore, it is not uncommon for a tumour to appear benign in some regions but have cancerous or pre-cancerous appearing regions [87]. Seeing as the detection algorithm would be used to suggest points of interest to a reading radiologist as part of a CAD tool or serve as an initialization for a renal mass classification algorithm, suggesting multiple areas of interest within a mass may be considered a useful feature in future applications.

5.1.4 Evaluation of Renal Mass Detection Accuracy

Similar to evaluation methods for the segmentation task, detection evaluation is completed in 3D. The primary method of detection evaluation was the computation of mass-wise precision and recall, defined in Eq. (11) and Eq. (12) respectively, based on the overlap method described in the previous section. Evaluating on a mass-wise basis as opposed to a slice-wise or patient-wise basis provides a more direct assessment of the models ability to identify renal masses. Accurately segmenting the boundaries of renal masses was not a priority for this study. The strength of multiparametric MRI based assessment of renal masses lies in its ability to classify based on pathology rather than determine morphological information, which is better suited to CECT imaging [15]. The DSC scores of renal mass segmentations were measured primarily for the purpose of understanding and comparing the strengths of different segmentation models.
The priority of renal mass detection is placed on recall because missing a suspicious region often has worse consequences than incorrectly labelling a healthy region as unhealthy. This decision was made based on the context of how a CAD system may theoretically be used clinically. If the detection algorithm were to be used along side a radiologist, its main purpose would be to catch suspicious features they may have missed, thereby reducing the false negatives of the radiologists readings [26]. Each component within the ensemble was individually optimized to maximize recall. The ensemble methods were also compared based on recall scores. While recall was the main priority, it was also important that the final model had high precision. A precise model would only suggest relevant regions of suspicion to a radiologist, making for a more useful tool.

The voxel-overlap based evaluation techniques only required an overlap of one voxel to register a mass as a true positive. This small overlap value was chosen to maximize recall, however it results in the precision and recall scores being insensitive to small localization errors in detections. To provide more context as to how well the algorithm finds the exact location of each renal mass, the centroids of the detected and ground truth masses are compared by calculating the average Euclidean centroid distance (CD) between them. Note that DSC was not used to assess detection localization accuracy. For the purpose of a CAD system, coarse localization based on the centroids was deemed to be more relevant than finding fine-grain voxel-wise accuracy.

5.2 Results

5.2.1 Renal Mass Segmentation

The RM segmentation algorithm was developed to maximize the recall metric for RM detection when used independently. The first test determined which method of localization to use. It revealed that background-elimination is superior to spatial-prior localization in terms of RM detection accuracy when using the ground truth kidney segmentations as shown in Table 11. Background-elimination localization was therefore used to determine optimal hyperparameters for RM segmentation. Notably, it was later found that spatial-prior localization outperforms background elimination when model-generated kidney segmentations were used. Since both techniques have similar architectures and target the same task, the hyperparameter tuning
completed in the following experiments for the background-elimination approach is expected to be applicable to the spatial-prior method as well.

Four major architecture changes were considered for renal mass segmentation including 3D and 2D variations of the standard and Attention U-Nets. The result of this testing is shown in Figure 21. On the basis of RM detection recall, it was clear that the 2D U-Net was the best performing model. Large standard deviations are seen in DSC and average centroid distance metrics for all models, especially compared to standard deviations seen for kidney segmentation in Chapter 4. This difference may be attributed to a greater degree of variation in morphology and intensities between renal masses making the segmentation task more challenging. As per Table 3, the volume of renal masses can range from approximately 2 cm$^3$ to 13 cm$^3$.

<table>
<thead>
<tr>
<th>Localization Method</th>
<th>DSC (%)</th>
<th>Detection Precision (%)</th>
<th>Detection Recall (%)</th>
<th>CD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background elimination</td>
<td>33.57 ± 30.96</td>
<td>45.7</td>
<td>76.2</td>
<td>15.82 ± 25.24</td>
</tr>
<tr>
<td>Spatial prior</td>
<td>37.52 ± 28.45</td>
<td>38.2</td>
<td>61.9</td>
<td>19.67 ± 27.57</td>
</tr>
</tbody>
</table>

Table 11: Renal mass segmentation and detection accuracy on the validation set using different kidney localization techniques. Results reported on the NG-MRI validation set using manual kidney segmentations for localization.
The next set of tests considered what model depth and loss function will deliver the highest RM detection recall on the validation set. Initially, a 6 layer model was used since it achieved the greatest kidney segmentation accuracy. This depth setting was verified by testing depths of 5 and 7 on the validation set. As per the results shown in Table 12, the 6-layer model reported the highest recall for renal mass detection. Notably, the 5 and 7 layer models both reported slightly better renal mass segmentation scores. However, since the focus of this study was on maximizing the recall metric for RM detection, the 6-layer approach was chosen going forward.

The loss function testing was completed using 5-fold cross validation on the training and validation set combined. $F_\beta$-BCE loss was considered here in addition to Dice loss, BCE loss and Dice-BCE loss. Attempting training with $\beta$ values above 1 was expected to increase the recall scores by penalizing false negative errors more than false positive errors. This trend was confirmed by the results shown in Table 13, where the model trained with a $F_\beta$-BCE loss and a $\beta$ value of 2.5 reported the highest renal mass detection recall. In fact, all the models trained with the asymmetric loss function reported a greater renal mass detection recall score than the Dice loss, BCE loss, and Dice-BCE loss.
Table 12: Renal mass segmentation and detection accuracy based on model depth. Results reported on the NG-MRI validation set using manual kidney segmentations for localization.

<table>
<thead>
<tr>
<th>Layer depth</th>
<th>DSC (%)</th>
<th>Detection Precision (%)</th>
<th>Detection Recall (%)</th>
<th>CD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Layers</td>
<td>35.98 ± 28.21</td>
<td>32.6</td>
<td>66.7</td>
<td>19.16 ± 28.81</td>
</tr>
<tr>
<td>6 Layers</td>
<td>33.57 ± 30.96</td>
<td>45.7</td>
<td>76.2</td>
<td>15.82 ± 25.24</td>
</tr>
<tr>
<td>7 Layers</td>
<td>37.15 ± 29.68</td>
<td>40.6</td>
<td>61.9</td>
<td>11.32 ± 11.03</td>
</tr>
</tbody>
</table>

Table 13: Renal mass segmentation and detection accuracy based on different loss functions. Results reported on the NG-MRI training and validation sets through 5-fold cross validation and using manual kidney segmentations for localization.

<table>
<thead>
<tr>
<th>Loss function</th>
<th>DSC (%)</th>
<th>Detection Precision (%)</th>
<th>Detection Recall (%)</th>
<th>CD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dice</td>
<td>33.16 ± 22.92</td>
<td>13.5</td>
<td>74.4</td>
<td>13.99 ± 17.17</td>
</tr>
<tr>
<td>BCE</td>
<td>36.65 ± 24.56</td>
<td>50.5</td>
<td>71.2</td>
<td>8.20 ± 9.58</td>
</tr>
<tr>
<td>Dice-BCE</td>
<td>44.46 ± 27.00</td>
<td>56.2</td>
<td>77.8</td>
<td>16.99 ± 19.69</td>
</tr>
<tr>
<td>$F_\beta$-BCE, $\beta=1.5$</td>
<td>44.91 ± 28.69</td>
<td>37.5</td>
<td>76.7</td>
<td>13.17 ± 15.86</td>
</tr>
<tr>
<td>$F_\beta$-BCE, $\beta=2.0$</td>
<td>47.04 ± 26.87</td>
<td>44.1</td>
<td>79.9</td>
<td>11.47 ± 11.54</td>
</tr>
<tr>
<td>$F_\beta$-BCE, $\beta=2.5$</td>
<td><strong>47.76 ± 25.31</strong></td>
<td>34.6</td>
<td><strong>83.8</strong></td>
<td>13.36 ± 15.33</td>
</tr>
</tbody>
</table>

To summarize prior results in RM segmentation, the optimal model was found to be a 2D U-Net with a depth of 6 layers and a $F_\beta$-BCE loss function with a $\beta$ value of 2.5. Lastly, consideration was given to assess if transfer learning from CECT data may improve renal mass segmentation and detection results. In the case of kidney segmentation in Chapter 4, no improvement was observed when the entire NG-MRI training set was used for fine-tuning the model. This test was performed to confirm if cross-modality training is beneficial for the renal mass detection task. To initialize the transfer learning model, it had to be pre-trained on the CECT data. This presented the opportunity to evaluate RM detection accuracy on the CECT test set, shown in Table 14.

While evaluation on CECT data was not an objective of this study, doing so highlights the challenge of working with the NG-MRI dataset. Without any modality specific hyperparameter
tuning, the model achieved a RM segmentation DSC of 78.51% ± 14.18% and a detection recall and precision in the 80% range on the CECT dataset. Despite the significantly greater performance on the CECT data, using the CECT data to pre-train the model for the task on the NG-MRI dataset did not result in any noticeable improvement. As per Table 15, a randomly initialized model achieved nearly identical results compared to the pre-trained model based on every evaluation metric. Transfer learning was therefore not included at this stage. Furthermore, it was decided that transfer learning would not be considered for the healthy tissue segmentation task and the ensemble approaches as well. The healthy tissue task is the complement of the renal mass segmentation task, and a component of the kidney segmentation task. Based on its similarity to both of these tasks, and the negligible improvement seen from transfer learning, it is highly likely that cross-modality training will not add any value to the final model.

Table 14: CECT Renal mass segmentation and detection accuracy. Results evaluated on the CECT test set using manual kidney segmentations for localization.

<table>
<thead>
<tr>
<th>Model</th>
<th>DSC (%)</th>
<th>Detection Precision (%)</th>
<th>Detection Recall (%)</th>
<th>CD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D U-Net with 6 Layers</td>
<td>78.51 ± 14.18</td>
<td>83.7</td>
<td>81.8</td>
<td>22.07 ± 21.47</td>
</tr>
</tbody>
</table>

Table 15: Effect of pre-training with CECT data on renal mass segmentation and detection performance. Results are reported on the NG-MRI validation set using manual kidney segmentations for localization.

<table>
<thead>
<tr>
<th>Model</th>
<th>DSC (%)</th>
<th>Detection Precision (%)</th>
<th>Detection Recall (%)</th>
<th>CD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without TFL</td>
<td>47.76 ± 25.31</td>
<td>34.6</td>
<td>84.4</td>
<td>13.36 ± 15.33</td>
</tr>
<tr>
<td>With TFL</td>
<td>45.29 ± 28.46</td>
<td>34.0</td>
<td>84.8</td>
<td>13.41 ± 13.58</td>
</tr>
</tbody>
</table>

5.2.2 Healthy Tissue Segmentation

Segmenting healthy tissue represents an alternate strategy for detecting renal masses, since any region that does not contain healthy tissue can be assumed to be a mass. In the next set of tests, an optimal healthy tissue segmentation model is found on the basis of renal mass detection accuracy. This section does not consider ensemble methods, but rather focuses on achieving the best results possible with healthy segmentation alone. To use the predicted healthy tissue
segmentation independently for RM detection it is converted to an RM segmentation by subtracting one and taking the absolute value of it. Based on the previously completed testing for RM segmentation and kidney segmentation, the architecture for healthy tissue segmentation was narrowed down to be a 2D U-Net with a depth of 6 layers. Transfer learning use was not considered for healthy tissue segmentation.

Both localization methods, background-elimination and spatial-priors, were independently considered for this task. As per the results shown in Table 16 below, background-elimination leads to the highest detection recall when using ground truth kidney segmentations. This result echoes localization testing for renal mass segmentation in Table 11. It appears that the background-elimination technique simplifies the task considerably, leading to better detection accuracy. This trend does not continue when working with model-generated kidney segmentations, as will be discussed in the next section. However, because of the similarity in the task and architecture between the two localization techniques, they use the same hyperparameters. Testing related to finding the optimal loss function for healthy tissue segmentation is therefore relevant regardless of which localization method is used. From the results in Table 16, it can also be observed that the healthy tissue segmentation models localization accuracy for renal masses is generally lower than the renal mass segmentation algorithm from the previous section. This may be because the RM segmentation model is directly trained to delineate RM boundaries, a trend that could be analogous to the use of a Dice loss leading to a better DSC score than using BCE loss as described earlier.

Dice, BCE, Dice-BCE, and Fβ-BCE loss functions were considered for this task. Due to healthy tissue segmentation being the complement to the renal mass segmentation task, precision was prioritized when training with the Fβ-BCE loss function by testing β values less than 1. Tuning the healthy tissue segmentation model for high precision is expected to have the effect of improving the recall metric for RM detection and is confirmed in results shown in Table 17. A β value of 0.25 was ultimately chosen for the final healthy tissue segmentation model, although it appears that the exact choice of β value has minimal impact on detection performance. This can be observed not only in the detection metrics, but also by the fact that all Fβ-BCE settings tested in Table 17 have nearly identical renal mass segmentation DSC scores.
Table 16: Effect of localization method on healthy tissue segmentation and RM detection accuracy. Results reported on the NG-MRI validation set using manual kidney segmentations for localization.

<table>
<thead>
<tr>
<th>Localization Method</th>
<th>DSC (%)</th>
<th>Detection Precision (%)</th>
<th>Detection Recall (%)</th>
<th>CD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background elimination</td>
<td>28.14 ± 29.81</td>
<td>44.7</td>
<td>81.0</td>
<td>47.37 ± 20.85</td>
</tr>
<tr>
<td>Spatial prior</td>
<td>11.67 ± 15.73</td>
<td>25.9</td>
<td>66.7</td>
<td>62.12 ± 25.97</td>
</tr>
</tbody>
</table>

Table 17: Loss function optimization for the healthy tissue segmentation model based on RM detection recall. Results reported on the NG-MRI training and validation sets through 5-fold cross validation using manual kidney segmentations for localization.

<table>
<thead>
<tr>
<th>Loss function</th>
<th>DSC (%)</th>
<th>Detection Precision (%)</th>
<th>Detection Recall (%)</th>
<th>CD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dice</td>
<td>25.10 ± 25.03</td>
<td>51.5</td>
<td>71.2</td>
<td>62.69 ± 22.19</td>
</tr>
<tr>
<td>BCE</td>
<td>25.09 ± 23.99</td>
<td>46.5</td>
<td>69.7</td>
<td>64.41 ± 22.19</td>
</tr>
<tr>
<td>Dice-BCE</td>
<td>27.97 ± 27.03</td>
<td>40.3</td>
<td>76.8</td>
<td>63.05 ± 21.49</td>
</tr>
<tr>
<td>F_{β}-BCE, β=0.5</td>
<td>36.95 ± 26.55</td>
<td>59.3</td>
<td>80.8</td>
<td>61.43 ± 22.61</td>
</tr>
<tr>
<td>F_{β}-BCE, β=0.25</td>
<td>36.30 ± 28.93</td>
<td>53.3</td>
<td>82.8</td>
<td>63.02 ± 22.58</td>
</tr>
<tr>
<td>F_{β}-BCE, β=0.125</td>
<td>37.70 ± 27.61</td>
<td>58.1</td>
<td>81.3</td>
<td>62.96 ± 21.57</td>
</tr>
</tbody>
</table>

5.2.3 Ensemble Models using Model Generated Kidney Segmentations

The previous two sections developed renal mass segmentation and healthy tissue segmentation models and evaluated their RM detection accuracy using ground truth kidney segmentations for localization. Results presented in this section evaluate RM detection accuracy when using kidney boundaries generated from the segmentation algorithm described in Chapter 4 for localization. Usage of each component independently for RM detection is compared with the two ensemble approaches, healthy tissue gating and mean averaging. Furthermore, since localization techniques may respond differently to kidney segmentation errors, both background-elimination and spatial-prior localization are considered here. To accomplish this, new models that accept the spatial-prior’s 2-channel input were trained. These models used the same hyperparameters as their equivalent background-elimination models, including loss functions. Evaluation of detection accuracy at this stage was completed on all 118 NG-MRI volumes through 5-fold cross validation.
validation. While initially the test set was considered for completing final evaluation, its sample size of just 12 patients found to lead to an over-representation of RM detection capability, so a cross validation approach was adopted.

Comparing the renal mass detection results in Table 18 to those in Table 19, it is clear that spatial-prior localization outperforms background-elimination localization in terms of RM detection when testing on fully autonomous data. All ensemble methods, including using each component independently, produced superior RM detection precision and recall scores using spatial-prior localization compared to their background-eliminated counterparts. Amongst all the ensemble methods tested with spatial-prior localization, the mean average ensemble produced the highest detection recall score of 86.2%. This technique also had a high precision score of 83.3%. These two metrics made it apparent that using mean averaging ensemble with spatial-prior localization leads to the most accurate RM detection results. In terms of RM localization accuracy, this method reported an average centroid distance error of 63.12 ± 28.15mm. This localization accuracy is worse than the mean-averaging ensemble model trained with background-elimination localization, which reported an average centroid distance error of 54.18 ± 25.91mm. The solid tumour-specific detection recall of the model was found in Table 20 to be 88.5%. This value is comparable to the detection recall for all renal masses, indicating that the detection capability for the final model is similar for fluid-filled cysts and solid tumours.

Table 18: Renal mass detection accuracy based on ensemble method when using background-elimination localization. Results reported on 118 NG-MRI volumes through 5-fold cross validation using model-generated kidney segmentations for localization.

<table>
<thead>
<tr>
<th>Ensemble Method</th>
<th>Renal Mass DSC (%)</th>
<th>Detection Precision (%)</th>
<th>Detection Recall (%)</th>
<th>CD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Averaging</td>
<td>28.49 ± 23.15</td>
<td><strong>56.9</strong></td>
<td>60.5</td>
<td>54.18 ± 25.91</td>
</tr>
<tr>
<td>Healthy Tissue Gating</td>
<td>19.83±20.24</td>
<td>47.4</td>
<td>45.2</td>
<td>63.79 ± 19.32</td>
</tr>
<tr>
<td>Renal mass segmentation only</td>
<td><strong>32.22 ± 22.62</strong></td>
<td>29.5</td>
<td><strong>75.7</strong></td>
<td><strong>12.66 ± 16.72</strong></td>
</tr>
</tbody>
</table>
Healthy tissue segmentation only

<table>
<thead>
<tr>
<th>Ensemble Method</th>
<th>Renal Mass Detection DSC (%)</th>
<th>Detection Precision (%)</th>
<th>Detection Recall (%)</th>
<th>CD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Averaging</td>
<td>25.71 ± 22.31</td>
<td>83.3</td>
<td>86.2</td>
<td>63.12 ± 28.15</td>
</tr>
<tr>
<td>Healthy Tissue Gating</td>
<td>14.80 ± 18.65</td>
<td>91.9</td>
<td>71.6</td>
<td>66.05 ± 27.09</td>
</tr>
<tr>
<td>Renal mass segmentation only</td>
<td>30.71 ± 20.79</td>
<td>34.5</td>
<td>76.7</td>
<td>16.41 ± 21.29</td>
</tr>
<tr>
<td>Healthy tissue segmentation only</td>
<td>14.02 ± 16.31</td>
<td>49.9</td>
<td>78.1</td>
<td>63.92 ± 26.51</td>
</tr>
</tbody>
</table>

Table 19: Renal mass detection accuracy based on ensemble method when using spatial-prior localization. Results reported on 118 NG-MRI volumes through 5-fold cross validation using model-generated kidney segmentations for localization.

<table>
<thead>
<tr>
<th>Model</th>
<th>Detection Recall (%)</th>
<th>CD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble RM Detector</td>
<td>88.5</td>
<td>63.51 ± 24.69</td>
</tr>
</tbody>
</table>

Table 20: Tumour specific detection recall and localization of the final RM detection model. Results reported on 118 NG-MRI volumes through 5-fold cross validation using model-generated kidney segmentations for localization.

Studies attempting RM detection often report results on a patient-wise basis. In [61], a patient-wise recall of 100% was reported, and in [65] a recall of 97.4% was reported on CECT datasets. While these results are not directly comparable to this study because of the modality, patient-wise detection capability of the final model is reported in Table 21 for context. A patient-wise recall of 90.0% was observed. For this purpose, a true positive was counted when at least one RM was successfully detected in an NG-MRI volume. Finding patient-wise precision was not possible since there were no patients that did not have any renal masses in the NG-MRI dataset. Further analyzing the patient-wise detection recall for solid renal masses in Table 22, it can be
seen that 89% of the 118 patients have at least 1 solid tumour detected. The patient-wise detection recall for benign tumours is lower than that of malignant tumours, indicating that there may be a bias to under-detect benign solid tumours.

Additionally, the ROC curve for the final renal mass detection model was plotted in Figure 22 to observe the RM segmentation quality. From this figure, it can be observed that the segmentation quality is not ideal. After a true positive rate of approximately 60%, the false positive rate accelerates quickly. Furthermore, the AUC value of this model was 0.718, significantly worse than the 0.972 AUC calculated for the kidney segmentation model proposed in Chapter 4.

Table 21: Evaluation of RM detection on patient-wise. Results reported on entire NG-MRI dataset using 5-fold cross validation using model-generated kidney segmentations for localization.

<table>
<thead>
<tr>
<th>Evaluation Method</th>
<th>Detection Precision (%)</th>
<th>Detection Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient-wise</td>
<td>N/A</td>
<td>90.0</td>
</tr>
</tbody>
</table>

Table 22: Patient-wise solid renal mass detection recall. Results reported on the entire NG-MRI dataset using 5-fold cross validation using model-generated kidney segmentations for localization.

<table>
<thead>
<tr>
<th>Renal Mass Type</th>
<th>Renal Mass Subtype</th>
<th>Detection Precision (%)</th>
<th>Detection Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solid Tumour</td>
<td>N/A</td>
<td>89.0</td>
<td></td>
</tr>
<tr>
<td>Malignant</td>
<td>N/A</td>
<td>91.8</td>
<td></td>
</tr>
<tr>
<td>Benign</td>
<td>N/A</td>
<td>81.8</td>
<td></td>
</tr>
</tbody>
</table>
In addition to quantitively analyzing the RM detection and segmentation results, the predicted RM boundaries were plotted in 2D and 3D, where further observations are made. Firstly, upon viewing the 2D axial slices in Figure 23, it can be observed that the RM segmentation produced by the model often does not follow the exact shape of the ground truth RM boundary. This is reflective of the relatively poor DSC score of the final model, 25.71%. When a successful detection is made by the model, it is the result of a specific feature or region within a RM being segmented. As seen in Figure 23 a), a cc-RCC tumour is detected in one axial slice but missed in a similar one. In this case, while the organ and surrounding tissue appear relatively similar between slices, the intensities and texture within the tumour show significant changes. This is reflective of the non-homogenous nature of malignant tumours and may be the reason masses are detected in some regions but missed in others. In the case of Figure 23 b) a cc-RCC tumour is successfully detected twice. The detections show significant differences in terms of intensity and texture, which may explain why they are separated. Figure 23 d) highlights a case where the model’s ability to detect spatially separated masses is a strength. This patient had a cyst and a solid tumour. In the left image, the model detected the cyst, and in the right, it detected the tumour. Both detected blobs could be analyzed separately by a future classification model, or a reading radiologist. These images also show that the model has a preference for segmenting
hypointense regions, such as in the case of Figure 23 d) where the cc-RCC is only detected in the darker center. This tendency may contribute to the poor localization accuracy of the model. Although, observing that the surrounding healthy tissue is generally hyperintense because of contrast perfusion, it is conceivable that hypointense regions would be less challenging to segment. An exception to this is the renal hilum, a healthy region which is often hypointense. In Figure 23 f) it can be seen that the papillary-RCC is successfully detected, but a false positive detection is made in a hypointense region of the renal hilum.

In Figure 24 below, RM detection results are visualized in 3D. While this figure further highlights the non-ideal RM segmentation of the model, it also provides context on the detection related strength of the model. Particularly, in the case on the right of Figure 24 where the patient has three separate RMs. The model was able to detect all three separately, indicated by the matching colors of the 3D blobs. The yellow and blue colored blobs in this figure also provide examples of false positive and false negative RM detections, respectively. The end-to-end computation time per patient, including kidney segmentation, ensemble SRM segmentation, post-processing, and SRM detection is 6.88 ± 0.52 minutes (mean ± SD).
Figure 23: a) – f) Predicted renal mass boundaries (red) are compared to ground truth renal mass boundaries (green) for patients with various renal mass pathologies. Two axial slices from each patient are shown.
Figure 24: 3D views of renal mass segmentations with matching detections highlighted by colour. Each column corresponds to a different patient. Segmentations generated by the deep learning model in the bottom row are compared to the ground truth segmentations in the top row.

5.3 Discussion

Prior to this study, detecting renal masses using MRI data has not been attempted, making the work presented in this chapter novel in its application. The described algorithm is capable of autonomously identifying spatially separated instances of renal masses in whole abdominal NG-MRI volumes. The final method for detecting renal masses was an ensemble containing two U-Net deep learning models, one to segment renal masses and another to segment healthy kidney
The models for both components have identical architectures with the exception of the $\beta$ value used in the F$\beta$-BCE loss function, but are trained to specialize in different subtasks in a Mixture-of Experts (MoE) style ensemble approach [62]. Both U-Net models had a depth of 6 layers and generated outputs on 2D axial slices. Input images for the renal mass and healthy tissue segmentation models were localized to the kidney region using kidney boundaries generated by the 2D U-Net model described in Chapter 4. Detection capability was evaluated on a dataset of 118 NG-MRI volumes collected from multiple healthcare institutions and from machines of different manufacturers as per Table 2. On the basis of mass-wise voxel overlap, the final model reported a precision and recall score of 83.3% and 86.2% respectively. Renal masses for this purpose contained both cysts and solid tumours. In terms of tumour-specific performance, the model reported a recall of 88.5%. On a patient-wise level, the model detected a renal mass in 90.0% of patients, making it well suited to the task of flagging abdominal NG-MRI volumes that contain renal masses. In terms of RM localization capability, the model reported an average centroid distance error of 6.31 ± 2.82cm. While this result is not ideal considering that the average diameter of a renal mass in the NG-MRI dataset is 3.34 ± 1.79cm, the requirement for spatial overlap by at least one voxel implies that all positive detections have a spatial relationship with a renal mass. Requiring overlap of more than one voxel would likely improve the localization results, but would be at the expense of a reduction in detection recall. Since maximizing the recall metric for RM detection is a greater priority than localization error, the one voxel overlap is still recommended. Localization accuracy may also be improved further by improving the healthy tissue segmentation component of the ensemble. It appears that it is the primary contributor to the loss of localization considering that the renal mass segmentation algorithm, when used alone, reports an average CD of 1.64 ± 2.13cm.

Previous studies related to renal mass detection used CECT images, making direct comparison to their results difficult. However, a unique strength of this study was its ability to detect multiple instances of renal masses in 3D and its evaluation on a mass-wise basis rather than a patient-wise or slice-wise basis, which has not been attempted using CECT images either. The work of Lin et al. in [65] which used U-Net models to autonomously detect RMs in CECT images, evaluated detection recall on a patient-wise level based on volume only. The work of Zabihollahy et al. in [61], which also performed autonomous RM detection in CECT images, evaluated detection capability based on 2D slice-wise pixel-overlap. The KiTS19 challenge avoided the RM
detection task and only focused on RM segmentation evaluation using DSC [51]. In comparison to all previous works related to autonomous RM detection, the evaluation method used in this study is an improvement because it measures the detection of a 3D object in 3D and incorporates spatial information.

The deep learning components of the proposed 2D U-Net ensemble model for RM detection were optimized for detection recall. Testing of different architectures and hyperparameter configurations related to model selection was completed for the renal mass segmentation algorithm. It revealed that the 2D U-Net outperformed its 3D and Attention U-Net counterparts. Furthermore, the 6-layer configuration outperformed the 5 and 7-layer configurations in terms of RM detection recall. These results echo testing in Chapter 4 for kidney segmentation, which also found a 2D U-Net with a depth of 6-layers to be optimal. This indicates that the tasks of kidney segmentation and renal mass detection are likely similar enough that similar hyperparameters can be used for both. A unique component of the models for RM and healthy tissue segmentation was the use of the Fβ-BCE loss. As opposed to kidney segmentation where precision and recall were equally important, here recall was given more importance. The Fβ-BCE loss function, which is an extension of the Fβ loss proposed in [49], successfully encouraged the model to prioritize recall during training as per the results in Table 13 and Table 17, where the models that reported the highest detection recall used Fβ-BCE loss implementations.

The optimal localization method for isolating the kidney region was found to be a spatial-prior method, where the kidney boundary is used to establish a 3D boundary around the organ and sliced axially. These axial slices were complemented by a second input channel that contained the binary kidney segmentation. This method outperformed the background-elimination localization technique when using model-generated kidney boundaries. It was less sensitive to under-segmentation errors in the kidney boundary since these image intensities were often still retained in the background of the input images. While under-segmented regions would be removed from the spatial-prior channel, the model has the opportunity to correct for these errors by still having access to the background. In contrast, the background-elimination method completely removes under-segmented regions, leaving no opportunity for compensation. The spatial-prior method does not completely eliminate the risk of under-segmentation errors. Such errors may still occur if an under-segmented region stretches beyond the coordinates for the 3D bounding box used to isolate the organ. However, upon observation the under-segmentation
errors by the kidney segmentation algorithm were very local in 3D and typically did not stretch beyond the bounding box. Furthermore, the length of a kidney is typically on an angle compared to the axial imaging plane. Since the bounding box is chosen to be aligned with the imaging planes, the bounding box is naturally wider than the kidney. Notably, the background-elimination technique outperformed spatial-priors when working with the ground truth kidney boundaries, as seen in the results in Table 11 and Table 16. This may be because the primary weakness of the background-elimination technique, under-segmentation errors, does not occur when ground truth kidney boundaries are used for localization. The background-elimination technique also produced superior RM localization results in terms of average CD error. A potential explanation for this may be that while background-eliminated models are more susceptible to under-segmentation errors, their feature extractors are more specialized since they do not need to learn to separate the kidney from the background. This localization technique may therefore reduce the impact of noisy elements in close proximity to the kidney, such as the liver and surrounding vasculature.

The ensemble method of mean-averaging was chosen over the other ensemble method, healthy gating, and the use of each component alone based on its superior RM detection recall and precision when using spatial-prior localization. While the healthy gating ensemble method did improve RM detection precision, it did not improve RM detection recall. This can be explained by how the healthy tissue segmentation model output is used by healthy gating ensembles. It is used to subtract regions from the renal mass segmentation that may be false positives but is not given the opportunity to contribute new segmented regions. As a result, it can improve on the precision of the renal mass segmentation model, but not the recall. In contrast the mean-averaging ensemble method allows both components to contribute to new segmented regions. The MoE ensemble approaches attempted in this study improves RM detection results by training models of the same architecture on different subtasks associated to the main task. The motivation for using this ensemble technique was that feature extractors used to output positive values for healthy kidney tissue may be easier to train than those designed to output positive values for all types of renal masses. Both sets of feature extractors will have different losses during training, since the areas they must cover are different, which may result in different types of errors and strengths at test time.
Another ensemble approach is to train models of different architectures on the same task, and then combine their outputs. This is a common approach, as seen in [61] for RM detection in CECT images where it outperformed using a single model. While such ensemble approaches were not attempted in this study, they may be able to further improve results of the model proposed in this thesis. As per the chart in Figure 21, the 3D U-Net models showed improved RM localization based on the average centroid distance. Future work may combine outputs from the 2D and 3D models to further improve the RM localization capability of the proposed model. Furthermore, recent advancements in image segmentation architecture, such as the nnU-Net may be implemented which have the potential to outperform the proposed model [52]. The testing completed in this thesis primarily focused on developing a well optimized U-Net model based on the original architecture, which is widely regarding as a gold standard for deep learning based medical image segmentation, to establish a baseline for autonomous RM detection in MRI.

Based on best practices, this study also addressed the issue of overfitting by augmenting the data using elastic deformations, using dropout layers during training, and using early stopping based on performance on a validation set. The issue of data imbalance within the images was primarily addressed by the choice of loss function, which asymmetrically penalized the model for missing renal masses.

Apart from the limited testing of architectures and ensembles, this study is limited by the data it has available. Similar studies performing RM detection using CECT images typically have over 300 segmented patient volumes [65]. Having more data available may increase the diversity of renal mass data available for training and evaluation. Specifically, rare tumour pathologies such as chromophobe RCC and angiomyolipoma were not well represented, with only 2 and 12 cases available in the dataset, respectively. Additionally, to establish the usefulness of the proposed algorithm as a CAD tool, its RM detection capability must be compared directly to that of radiologists. With such data, we may see improved inter-operator variability and higher detection recall. Although, that type of assessment is not possible in this study since the available ground truth results used to evaluate model RM detection accuracy incorporates information from biopsy reports and all other MRI sequences available.

Since this data was collected retrospectively, it also must be acknowledged that there is potential for confirmation bias within these results [88]. This type of bias would only be relevant if the proposed algorithm were to be considered to as a screening tool on MR images of patients with
unknown renal masses. Confirmation bias may occur if the MR images in this dataset are primarily from patients that were referred for an MRI after it was already known that they have a challenging renal mass after CECT imaging, as opposed to patients that are merely suspected of having a renal mass because of incidental reasons. Renal masses in patients that are investigated with CECT and then referred for MR imaging may have characteristics that are easier to detect, making them non-ideal representations of the whole population that presents with incidentally suspected renal masses. This scenario is specifically relevant to this dataset, since CECT has historically been the first line of assessment for those suspected of having a renal mass, with MRI often reserved for those that have masses with unclear pathology or those that cannot tolerate CT imaging [3]. To establish the effectivity of the proposed model as a CAD tool, further testing using MR images of patients with suspected renal masses must be performed, since they may contain masses that are not normally indicative of needing an MRI investigation. Additionally, this study only included cases where biopsy or surgical confirmation of pathology is available. While this makes the data free of inter-operator variability from reading radiologists, it ignores potential cases where a benign mass was identified in an MRI and a biopsy was never completed. In such cases acquiring biopsy confirmation is not possible since it would not provide a clinical benefit to the patient given that the mass appears obviously benign on imaging. An alternate strategy to include such patients in the study may be to follow their clinical history several years after a benign mass diagnosis to confirm that it is in fact benign.

Despite these limitations, the results of this study are highly relevant to the clinical community as it increasingly considers incorporating MRI in renal mass assessments [14]. Future work in this area may use the proposed ensemble methods to define regions suspected to contain renal masses and perform classification tasks such as establishing tumour malignancy. In this pursuit, it is highly recommended to expand ensemble models to include the multi-parametric data available for each patient. Adding information from more sequences will likely improve RM detection accuracy further since masses of different pathology are affected differently by MRI sequences based on their underlying anatomy and physiology [21].

5.4 Conclusion

The proposed renal mass detection algorithm is able to accurately locate spatially separate instances of renal masses within NG-MRI data. An ensemble of 2D U-Nets is used to generate
segmentation maps of renal masses. These segmentation maps are compiled into 3D volumes where unique masses are identified. Evaluation based on mass-wise voxel overlap reported a final detection precision and recall score of 83.3% and 86.2% respectively on 118 NG-MRI volumes through 5-fold cross validation. Masses are localized well, with an average centroid distance between predicted and true mass locations being $6.31 \pm 2.82$cm. The proposed model provides RM detection with instance localization on a fully autonomous basis using entire abdominal NG-MRI volumes. Based on its high detection precision and recall as well as reasonable localization accuracy, the proposed model is well suited to be used as an initialization point for renal mass classification algorithms. Although further testing is needed with real world data, this algorithm may also be used as a standalone CAD tool to identify renal masses in patients and enable early and effective treatment.
Chapter 6: Conclusion and Recommendations

The work presented in this thesis describes a fully autonomous renal mass detection algorithm using NG-MRI images. This algorithm works in three distinct stages. Firstly, the boundary of the kidney is identified within whole abdominal MRI volumes using a kidney segmentation algorithm. Kidney boundaries are then used to localize a region of interest which contains renal masses. The next stage involves segmenting renal masses and healthy kidney tissue using two independent segmentation algorithms. Finally, an ensemble approach combines the renal mass and healthy tissue segmentations. All the presented segmentation algorithms followed a deep learning approach based on the 2D U-Net architecture. Boundaries were generated on axial 2D slices which were then compiled into 3D volumes for segmentation and detection accuracy evaluation.

The 2D U-Net with a depth of 6 layers was chosen over the 3D U-Net and Attention U-Net architectures, primarily due to similar segmentation accuracy while being smaller models in size with fewer tunable parameters. Deep learning models for the kidney segmentation, renal mass segmentation, and healthy tissue segmentation were similar, with their main difference being the choice of loss function and the number of input channels. Transfer learning by pre-training on CECT data was attempted but was found to be unnecessary for achieving high detection accuracy based on the size of the NG-MRI dataset available for this study. Although, future studies that are struggling with a very small dataset size may consider cross-modality transfer learning and gain a significant improvement in accuracy. For kidney segmentation, a combination of binary cross entropy (BCE) and Dice loss was found to produce more accurate boundaries than BCE or Dice loss alone. For renal mass and healthy tissue segmentation, the $F\beta$-BCE loss function played an important role by enabling the training process to prioritize high recall through a tunable hyperparameter. Another differentiating factor between the kidney segmentation and the renal mass and healthy tissue segmentation models was the use of a localization channel. While the kidney segmentation model operated on whole axial MRI slices, subsequent models operated on localized slices with a separate channel that contained the kidney segmentation boundary. This localization channel acted as a spatial prior which provided pixel-wise information regarding the kidney boundary while accommodating for under-segmentation errors by the kidney segmentation algorithm leading to higher detection recall.
Maximizing the recall metric for RM detection was a priority for this study since the clinical implications for a false negative are worse than a false positive detection. The ensemble approach, using mean averaging to combine renal mass and healthy tissue segmentations, was the most effective in terms of recall scores. It outperformed using the renal mass or healthy tissue segmentations alone for detection, which reported a recall of 77.7% and 78.0% respectively on the full NG-MRI dataset. Mean averaging ensemble reported a fully autonomous renal mass detection recall of 86.2% and a precision of 83.3% on the full NG-MRI dataset. This result highlights that the proposed CAD system is effective at autonomously detecting renal masses in NG-MRI abdominal volumes.

My work is relevant to the increased interest clinically in using MR imaging for renal mass assessment. It supports the use of deep learning based approaches to autonomously detect renal masses of interest within this modality. If used clinically by complimenting a radiologist reading, it has the potential to reduce the incidence of false negative readings where the radiologist misses a potentially diagnostic area in the image. Although, to make this assertion with certainty, detection results must be compared to the accuracy of radiologists for the same data. All available ground truth data in this study was influenced by biopsy and surgical reports in addition to radiologist assessments, making it impossible to compare deep learning results to the accuracy of an independent radiologist reading.

Additionally, significant clinical value of using MRI for renal mass assessment lies in the classification of renal masses using multiparametric imaging where different sequences highlight pathological differences [21]. It is therefore recommended to expand this study to include renal mass classification, where solid tumours can be separated from fluid-filled cysts and malignant masses can be separated from benign ones. Effectively classifying benign from malignant masses will help address another clinical issue, the overtreatment of patients with benign renal masses [15]. The autonomous detection algorithm I proposed can support classification studies by suggesting regions of interest that contain renal masses, making it possible to attempt fully autonomous classification. Furthermore, expanding the ensemble approach presented in my work to include co-registered multiparametric data from each patient may also improve detection results directly.
The final objective of this study was to develop a fully autonomous renal mass detection algorithm, with the motivation of encouraging the development of CAD tools for renal mass detection using MRI. While this task has been extensively explored using CT imaging [51], there have been no previous studies attempting renal mass detection using MRI. In this regard, the results in this thesis fulfil its intended objective and provide a significant contribution to the field of computer aided detection. Further studies directly comparing detection accuracy to CT and independent radiologist readings on the same patient data and attempting renal mass classification may further solidify the role of MRI for renal mass assessment.

For the task of fully autonomous renal mass detection in NG-MRI, I conclude that the proposed algorithm using a cascaded ensemble of 2D U-Nets is effective and accurate. It may serve as an initialization point for further studies in renal mass classification and improve outcomes for patients by assisting with the detection of potentially malignant regions in the kidney.
REFERENCES


[66] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, “Panoptic segmentation,” arXiv,


APPENDIX A: FIGURE REPRODUCTION PERMISSIONS

Permission for Figure 3:

Thesis / Dissertation Reuse
The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of University of Guelph's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.
Permission for Figure 4:

<table>
<thead>
<tr>
<th>SPRINGER NATURE LICENSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TERMS AND CONDITIONS</td>
</tr>
<tr>
<td>Jul 04, 2021</td>
</tr>
</tbody>
</table>

This Agreement between University of Guelph -- Mr. Anush Agarwal ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

<table>
<thead>
<tr>
<th>License Number</th>
<th>5102260209924</th>
</tr>
</thead>
<tbody>
<tr>
<td>License date</td>
<td>Jul 04, 2021</td>
</tr>
<tr>
<td>Licensed Content Publisher</td>
<td>Springer Nature</td>
</tr>
<tr>
<td>Licensed Content Publication</td>
<td>Springer eBook</td>
</tr>
<tr>
<td>Licensed Content Title</td>
<td>U-Net: Convolutional Networks for Biomedical Image Segmentation</td>
</tr>
<tr>
<td>Licensed Content Author</td>
<td>Olaf Ronneberger, Philipp Fischer, Thomas Brox</td>
</tr>
<tr>
<td>Licensed Content Date</td>
<td>Jan 1, 2015</td>
</tr>
<tr>
<td>Type of Use</td>
<td>Thesis/Dissertation</td>
</tr>
<tr>
<td>Requestor type</td>
<td>academic/university or research institute</td>
</tr>
<tr>
<td>Format</td>
<td>print and electronic</td>
</tr>
<tr>
<td>Portion</td>
<td>figures/tables/illustrations</td>
</tr>
<tr>
<td>Number of figures/tables/illustrations</td>
<td>1</td>
</tr>
<tr>
<td>Will you be translating?</td>
<td>no</td>
</tr>
<tr>
<td>Circulation/distribution</td>
<td>1 - 29</td>
</tr>
<tr>
<td>Author of this Springer Nature content</td>
<td>no</td>
</tr>
<tr>
<td>Title</td>
<td>Automated Detection of Renal Masses in Contrast-Enhanced MRI using Deep Learning Methods</td>
</tr>
<tr>
<td>Institution name</td>
<td>University of Guelph</td>
</tr>
<tr>
<td>Expected presentation date</td>
<td>Aug 2021</td>
</tr>
<tr>
<td>Portions</td>
<td>Figure 1: U-net architecture</td>
</tr>
</tbody>
</table>

103
Terms and Conditions

This agreement sets out the terms and conditions of the licence (the Licence) between you and Springer Nature Customer Service Centre GmbH (the Licensor). By clicking 'accept' and completing the transaction for the material (Licensed Material), you also confirm your acceptance of these terms and conditions.

1. Grant of License
   1. The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.
   2. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).
   3. If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Scope of Licence
   1. You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.
   2. A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights

104
have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

3. Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

4. Where permission has been granted free of charge for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

5. An alternative scope of licence may apply to signatories of the STM Permissions Guidelines, as amended from time to time.

- **Duration of Licence**

  1. A licence for is valid from the date of purchase ('Licence Date') at the end of the relevant period in the below table:

     | Scope of Licence       | Duration of Licence          |
     |------------------------|------------------------------|
     | Post on a website      | 12 months                    |
     | Presentations          | 12 months                    |
     | Books and journals     | Lifetime of the edition in the language purchased |

- **Acknowledgement**

  1. The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

- **Restrictions on use**

  1. Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that affect the meaning, intention or moral rights of the author are strictly prohibited.
2. You must not use any Licensed Material as part of any design or trademark.
3. Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

• Ownership of Rights

1. Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

• Warranty

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

• Limitations

1. **BOOKS ONLY:** Where 'reuse in a dissertation/thesis' has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline ([www.sherpa.ac.uk/romeo/](http://www.sherpa.ac.uk/romeo/)).
2. For content reuse requests that qualify for permission under the [STM Permissions Guidelines](http://www.sherpa.ac.uk/romeo/), which may be updated from time to time, the STM
Permissions Guidelines supersede the terms and conditions contained in this licence.

- **Termination and Cancellation**

1. Licences will expire after the period shown in Clause 3 (above).
2. Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

**Appendix 1 — Acknowledgements:**

**For Journal Content:**

Reprinted by permission from [the Licensor]: [Journal Publisher (e.g. Nature/Springer/Palgrave)] [JOURNAL NAME] [REFERENCE CITATION (Article name, Author(s) Name), [COPYRIGHT] (year of publication)]

For **Advance Online Publication papers:**

Reprinted by permission from [the Licensor]: [Journal Publisher (e.g. Nature/Springer/Palgrave)] [JOURNAL NAME] [REFERENCE CITATION (Article name, Author(s) Name), [COPYRIGHT] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)]

**For Adaptations/Translations:**

Adapted/Translated by permission from [the Licensor]: [Journal Publisher (e.g. Nature/Springer/Palgrave)] [JOURNAL NAME] [REFERENCE CITATION (Article name, Author(s) Name), [COPYRIGHT] (year of publication)]

**Note: For any republication from the British Journal of Cancer, the following credit line style applies:**

Reprinted/adapted/translated by permission from [the Licensor]: on behalf of Cancer Research UK: : [Journal Publisher (e.g. Nature/Springer/Palgrave)]