Improving Segmentation Pipelines for Medical Imaging using Deep Learning

by

Jay Biren Patel

B.S., Case Western Reserve University (2016)

Submitted to the Harvard-MIT Program in Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Medical Engineering and Medical Physics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by: Jay Biren Patel
Harvard-MIT Program in Health Sciences and Technology
May 11, 2023

Certified by: Jayashree Kalpathy-Cramer, PhD
Professor of Ophthalmology, Chief of Division of Artificial Medical Intelligence in Ophthalmology, University of Colorado School of Medicine
Visiting Professor in Radiology, Harvard Medical School
Thesis Supervisor

Accepted by: Collin M. Stultz, MD, PhD
Director, Harvard-MIT Program in Health Sciences and Technology
Nina T. and Robert H. Rubin Professor in Medical Engineering and Science
Professor of Electrical Engineering and Computer Science

Improving Segmentation Pipelines for Medical Imaging using Deep Learning

by

Jay Biren Patel

Submitted to the Harvard-MIT Program in Health Sciences and Technology
on May 11, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Medical Engineering and Medical Physics

Abstract

One of the most important steps in the clinical workflow is the segmentation of medical imaging, which can be used for a variety of clinical decision-making tasks such as disease diagnosis and treatment response evaluation. Manual segmentation of 3D medical imaging (such as computed tomography (CT) or magnetic resonance imaging (MRI)) by a clinical expert can be too time-consuming to be feasible in a routine clinical workflow, and can moreover be susceptible to human errors and inconsistencies. In recent years, deep learning (DL) based methods have exhibited human-level performance for a variety of computer vision tasks, making them an attractive choice for researchers aiming to automate the segmentation of medical imaging. This thesis considers two medical imaging scenarios and examines how fully automatic image segmentation via DL can enhance downstream clinical tasks.

The first scenario evaluates the clinical workflow for diagnosing incidental adrenal masses on CT. Despite standardized reporting systems and strict guidelines for defining an adrenal mass, there exists significant inter-rater variability for this task. To enable objective and reproducible characterization of the adrenal gland, this thesis develops the first DL method for segmentation and classification on CT. Using a large-scale retrospectively acquired dataset, this method is used to identify potential missed detections by radiologists and discuss the clinical implications of this.

The second scenario focuses on the treatment response assessment of metastatic brain tumor patients on MRI. Due to the large number of metastases a patient can have, standard radiographic analyses track only a select few target lesions through the course of therapy in order to assess the efficacy of a treatment. With this paradigm, smaller non-target lesions may be neglected or even missed due to the lack of quantitative emphasis. To that end, a pipeline is developed to automatically segment brain tumor metastases on MRI and output standard response assessment metrics. With the prevalence of longitudinal imaging data available for brain metastases patients, a secondary model is formulated to improve the detection and segmentation of micro-metastases by utilizing known prior time point information.

Thesis Supervisor: Jayashree Kalpathy-Cramer, PhD
Title: Professor of Ophthalmology, Chief of Division of Artificial Medical Intelligence in Ophthalmology, University of Colorado School of Medicine
Visiting Professor in Radiology, Harvard Medical School

Acknowledgments

"Try to learn something about everything and everything about something." When I was making my Facebook profile way back in middle school, I had proudly displayed this quotation by Thomas Henry Huxley in the about me section of my profile. You see, even before I had the faintest idea what type of major I would like to study in college, I sub-consciously always knew I would pursue a doctoral degree eventually.

Having just finished my thesis, I can confidently say that middle school me was not prepared for just how challenging it would be and just how much I would rely on others for support and guidance throughout the process. First and foremost, I need to thank my advisor Jayashree, whose guidance and mentorship has been invaluable throughout my PhD. I am deeply grateful for her patience and understanding and need to emphasize that she really allowed me to flourish throughout this process. Doing my PhD would not have been nearly as enriching without her.

Next, I need to thank the rest of my committee: Elizabeth, Bruce, and Elfar. Elizabeth has been an invaluable resource to me throughout my PhD. Without her, the clinical motivation for my work would be nowhere near as strong. I also need to acknowledge Bruce and Elfar. The insightful comments and feedback that was provided during committee meetings really helped to elevate the quality of the final thesis.

Next, I'd like to thank some of my collaborators and mentors. First, Dr. Randy Gollub for all her help and guidance near the beginning of my PhD. I also need to thank her for allowing me to be a part of the NTP group. Next, Dr. Bill Mayo-Smith, who really guided the adrenal segmentation and classification project for me. Over the course of my time in Jayashree's lab, I've worked on numerous projects, some of which did not get into the final thesis. One such project was stroke lesion segmentation with Dr. Natalia Rost (who helped fund part of my PhD). Another such project was adrenoleukodystrophy lesion segmentation, with Dr. Patty Musolino. I want to thank her for all her clinical expertise and knowledge she imparted to me.

I am immensely grateful to have such a wonderful lab. Whether its playing spikeball, or going climbing, or just getting lunch, I could not have asked for a better set of colleagues and friends. In particular, I need to thank Kevin Lou, Ikbeom Jang, Praveer Singh, Matthew Li, Mehak Aggarwal, Sharut Gupta, Albert Kim, Ben Bearce, Chris Bridge, Mishka Gidwani, Kathi Hoebel, Ken Chang, and Syed Rakin Ahmed.

Similarly, I am thankful that the HST cohort of mine is so close knit and special. Trivia nights and friendsgiving and ski trips have got me through the last few years. I am so lucky to have met friends like Melinda Chen, Lucy Hu, Erin Rousseau, Melodi Anahtar, and Gowtham Thakku.

I of course need to thank my partner Charu, who has been an incredible source of inspiration in my life. She always knows the right thing to say when I'm stressed and has been so kind and patient and compassionate these past few months while I have been finishing up.

Finally, I need to acknowledge my family. My dad and brother have been so loving and supportive these past two years, and I'm grateful that they could be here today to see me. I am thankful to my mom. I would not be the person I am today without her, and I know she would have been proud.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Medical Imaging

The first medical image ever acquired was taken on November 8$^{\text{th}}$ 1895, when Conrad Roentgen took the world's first x-ray of his wife's hand. This discovery was met with immediate amazement and wonder by medical professionals, who went on to publish over $1,000$ scientific articles on x-rays within the next year [1]. By 1956, advances in physics and electrical engineering led to the first ultrasound, and by the late 1970's, the first 3D images were being generated via computed tomography (CT) and magnetic resonance imaging (MRI). Since then, medical imaging has revolutionized the field of healthcare, enabling accurate and completely non-invasive diagnosis for a wide spectrum of diseases. It is thus not surprising that healthcare professionals are acquiring an exponentially growing number of medical images every year. Indeed, the world health organization (WHO) estimates that an approximate 3.6 billion examinations are now being performed worldwide per year [2].

Historically, the interpretation of medical imaging by radiologists and other clinicians has been mostly qualitative [3]. When given a new examination to interpret, the radiologist would identify abnormalities such as tumors, fractures, diseases, etc. based on their subjective assessment of the shape, size, and image intensity of different anatomical regions of interest (ROI). However, it is important to note that qualitative interpretation of imaging can be influenced by a variety of factors, including the expertise, personal experience, and expectations of the interpreter. This can lead to significant inter- and intra-rater variability between different clinicians and can potentially affect the accuracy of the diagnosis. One way to reduce the impact of bias in subjective interpretation is through the use of standardized reporting systems and guidelines, which can help to encourage that all clinicians are interpreting the images in a consistent and objective manner. However, even with firm guidelines in place, many studies still report significant amounts of inter-rater variability for certain tasks such as the interpretation of chest x-rays [4, 5, 6] and detection of primary and metastatic tumors [7, 8, 9].

1.2 Medical Image Segmentation

With a desire to make the interpretation of medical imaging more objective and reproducible, clinicians have begun to switch towards using a more quantitative approach. One of the most common ways to do this is via segmentation, wherein specific anatomical structures or abnormalities are delineated within the image [10]. The goal is to identify certain ROIs which can be used to guide disease diagnosis, aid in surgical planning, enhance treatment response evaluation, and/or help in other aspects of clinical decision-making. For example, a clinician may segment a patient's tumor on a pre-treatment and post-treatment MRI. If the tumor volume has not significantly shrunk after application of the treatment, the clinician may decide to switch treatment plans in order to find something that more efficacious.

With an ever-increasing number of high resolution imaging sequences being acquired, it can often be infeasible in a routine clinical workflow to manually segment all imaging. Moreover, even when manual segmentations are possible, there are a host of associated issues [11]. First, segmentation can be a challenging task that requires a high degree of expertise and precision [12]. While this may not be a problem at a major hospital system, the level of specialization required to interpret and segment all different types of imaging sequences may not be available for smaller, more rural clinical settings. Second, many anatomical structures can blend together on macroscopic imaging such as CT due to issues stemming from image resolution and image contrast, leading to subjectivity and variability in how different interpreters identify and segment certain ROIs. This issue can be exacerbated further due to poor image quality, which may result from things like motion and aliasing artifacts [13]. Finally, even when image quality is perfect, manual segmentation is susceptible to human errors and inconsistencies [14].

1.3 Deep Learning for Medical Imaging

To combat the issues surrounding manual image segmentation and to alleviate the workload for healthcare professionals, deep learning (DL) based solutions for image segmentation are becoming increasingly common [10, 15, 16]. Unlike classical machine learning (ML) approaches which require considerable feature engineering and domain expertise in order to design relevant hand-crafted imaging features, data-driven DL methods can learn salient features directly from the training data [17]. This data-driven approach has been proven fruitful, having reached (and sometimes surpassed) human level performance on a plethora of tasks including primary brain tumor segmentation [18].

With that being stated, there are still many challenges hindering the successful deployment and utilization of DL based models in the clinical workflow. In this thesis, we consider two medical imaging scenarios and examine how fully automatic image segmentation via DL can enhance associated downstream clinical tasks. Specifically, we consider the scenarios of 1) adrenal gland segmentation and classification on CT and 2) metastatic tumor segmentation on MRI.

1.4 Thesis Organization

This thesis is structured as follows:

Chapter 2 introduces prerequiste background material that is necessary to understanding the work presented in this thesis. We provide a brief overview of DL using convolutional neural networks, explaining relevant topics such as data augmentation and loss functions. We also provide summaries of selected classical and DL based methods for image segmentation and image registration.

In chapter 3, we present a novel two-stage DL based pipeline for adrenal gland segmentation and classification, which we validate on a real-world consecutively acquired dataset. We also present an exploratory analysis on the inter-rater variability between expert radiologists for the task of adrenal gland classification, as well as use our model to identify potential missed detections on a large-scale retrospective dataset.

In chapter 4, we train a DL model for segmentation of metastatic brain tumors and validate on two independent datasets. We also generate voxel-wise segmentation uncertainty maps, using them to automatically flag potential false positives. Using our model outputs, we automate longitudinal volumetric tracking of metastases and present an algorithm to automate current uni-dimensional response assessment criteria. Finally, we conclude with a short analysis about the differences between true volumetric tumor burden and proxy uni-dimensional measures.

In chapter 5, we aim to further improve our work on metastatic tumor segmentation. Namely, to improve the sensitivity of detection for micro-metastatic lesions, we develop a novel DL based approach for joint image registration and segmentation. As prior time-point imaging (and prior time-point segmentations) are readily available in a routine clinical setting, we devise a method which can incorporate this known prior information to improve the segmentation of the new time-point. Using our model, we show promising results, noting a significant increase in the detection rate of micro-metastatic lesions.

Chapter 2

Background

In this chapter, we review some necessary background information that will aid the reader in understanding the rest of this thesis. First, we provide a light overview of what a neural network is and some associated terminology (e.g. loss functions, optimization, overfitting, etc.) in sections 2.1 and 2.2. Second, we describe common image segmentation approaches in the context of medical imaging in section 2.3. Third, we provide a review of classical and DL based registration methodologies in section 2.4. Finally, we provide brief clinical insights into the tasks of adrenal gland segmentation in section 2.5 and brain metastases segmentation in section 2.6.

2.1 Neural Networks and Deep Learning

A neural network is a model that takes a raw image as input and applies many layers of learned transformations to calculate some desired output [19]. Each layer in the network is comprised of nodes, known as neurons. The value at each neuron is calculated by taking a linear combination of neurons in the previous layer, and applying some non-linearity known as the activation function [20]. Generally speaking, the representational power (or expressiveness) of neural networks comes from the repeated chaining of activation functions, which allow the network to learn complicated concepts by building them out of simpler ones [19]. This hierarchy of concepts forms the basis for deep learning, which states that as a network's depth (i.e. layers), width (i.e. neurons per layer), or connectivity (i.e. number of connections between neurons in different layers) increases, the network is able to approximate a larger family of functions, an axiom loosely encapsulated in the universal approximation theorem [21, 22].

There are three main ways to train a neural network: supervised learning, semi-supervised learning, and unsupervised learning. In supervised learning, we create paired input-output data (known as labeled data) [23]. What comprises these input-output pairs will change depending on the intended task. For example, for the task of image classification, input would refer to a sample image and output would refer to what ground truth category it was (e.g. cat, etc.). For the task of image segmentation, the input would once again refer to a sample image and output would be the ground truth segmentation (e.g. the cat delineated from the rest of the image). Once we have curated a paired dataset, we can train our network to learn directly from this cohort. It is important to note that paired data can be expensive and difficult to curate. To combat this issue, semi-supervised approaches can be used [24]. In this paradigm, large quantities of unlabeled input data can be used simultaneously with a small fraction of labeled cases to improve the quality of the output. Indeed, several studies have shown that leveraging this unlabeled data can improve model performance by non-trivial amounts [25, 26]. Finally, fully unsupervised (or sometimes self-supervised) approaches require no ground truth labels whatsoever. While this may simplify the task of data collection and curation, it can make learning relevant information more difficult, especially if one has a niche task. Nonetheless, unsupervised learning has been used to generate state-of-the-art results on image classification tasks [27, 28]. A diagram illustrating simple examples of each of the three learning paradigms is shown in figure 2-1.

While fully connected neural networks tend to perform better than what is capable via traditional machine learning techniques, it was the development of convolutional neural networks (CNN) that pushed the envelope of artificial intelligence to where it is today [29, 30]. CNNs, as the name implies, use kernels which are convolved (though in actuality they are cross-correlated) with an input signal. These kernels are usually

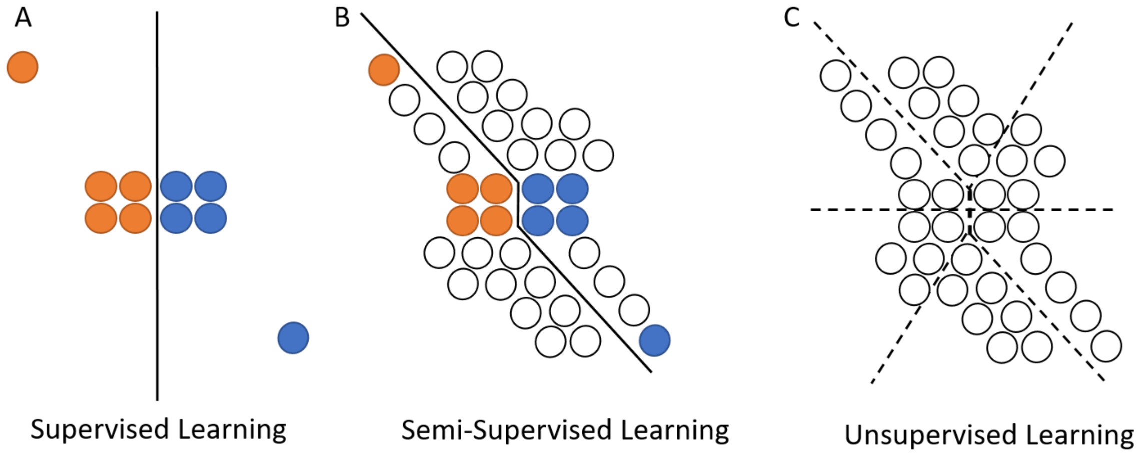


Figure 2-1: **Different types of learning paradigms.** A) Due to the difficulty in collecting large amounts of labeled data, fully supervised learning approaches tend to have only a limited number of labeled samples. B) By using unlabeled data, we can leverage semi-supervised approaches to find a more likely decision boundary. C) Unsupervised approaches can be used to find decision boundaries, but they may partition the dataset incorrectly or in an undesirable manner

chosen to be significantly smaller than the input signal, which means that the output at each point depends only on the neurons which are within the window size of the kernel. This local window size that the kernel can see is known as the receptive field (RF) of the kernel, and it is important because it determines the size and complexity of the features that the kernel can detect. Larger kernels have larger receptive fields, allowing them to capture more global features of the input image, while smaller kernels can capture more local features [31]. CNNs perform much better than fully connected networks on standard computer vision tasks because they are designed to exploit the spatial structure of images and capture local patterns and features between nearby pixels. Moreover, CNNs have been shown to learn very hierarchal representations. In the initial layers, learned kernels typically detect simple patterns such as edges and colors. These low-level features are combined in subsequent layers to form more complex patterns, such as shapes and textures. As the input image is processed through deeper layers of the network, the effective receptive field (ERF) of the network increases, allowing the kernels to capture increasingly more complex and abstract concepts [32].

Presently, CNNs exhibit state-of-the-art performance on a wide range of computer vision tasks such as image classification [33, 34, 35, 36], semantic segmentation [37, 38, 39], image enhancement [40, 41], etc. More recently, these methods have also been applied to the medical imaging sector, where they have been used successfully for tasks such as tumor segmentation [42, 43, 44, 45, 46], prediction of mutation status for gliomas [47], and automatic grading of gliomas [48].

2.2 Training Deep Learning Models for Medical Imaging

2.2.1 The Risk of Overfitting

When training a neural network, it is common practice to monitor both the training and validation performance. The inflection point where training performance still improves but validation performance starts to decrease is the beginning of model overfitting. For a neural network to generalize well to unseen data, it is generally agreed upon that the training set must be large and diverse in order to allow the neural network to effectively model the entire distribution [49]. This is especially important when there exists only subtle differences between imaging phenotypes, as one might see on non-contrast imaging, or if there exists significant heterogeneity in the input data, as one might see when utilizing multi-institutional datasets. Unfortunately, large quantities of high quality ground truth annotations for medical image segmentation tasks do not always exist, resulting in small sample sizes which will adversely affect the generalizability of the model [50, 51, 52].

Improving the generalizability of a neural network (which can be approximately framed as reducing the amount of overfitting on the training set) is a well-studied problem and many solutions have been proposed with varying degrees of success. The first class of solutions involves changes to the model itself. For example, overfitting can be mitigated by reducing the overall capacity of the model (which is accomplished by decreasing the depth, width, or connectivity of the network) [53]. Other techniques include the addition of dropout layers or explicit constraints on the kernel weights of the network through weight decay or L1/L2 regularization [54, 55, 56]. While these generic approaches can yield expressive networks that are satisfactorily generalizable, they are not necessarily optimal because they are not tailored towards any specific task or input dataset.

2.2.2 Data Augmentation

Data augmentation is the process of increasing the diversity of inputs seen by a neural network by generating randomly transformed versions of the given training set. Augmentation methods can be broadly stratified into two categories: spatial and intensity transformations. Spatial augmentations include random cropping, flipping, rotations, scaling, shearing, aspect ratio modifications, and elastic deformations; intensity augmentations include random shifts in brightness, contrast, saturation, and hue. Recent results have shown that correct use of data augmentation can significantly improve performance on both the training and validation sets [57, 58, 59]. However, data augmentation methods require expertise and manual work to design policies that capture prior knowledge in each domain. A simple example can be seen in digit recognition, where large rotations can potentially be non-label preserving (e.g. a 180° rotation of a "6" will create a "9", thus creating false training examples). Unsurprisingly, it follows that incorrect data augmentation may in fact lead to worse performance, as highly distorted/corrupted generated samples lower the signal-to-noise ratio and force the network to learn representations not indicative of the true data distribution [60]. To address this shortfall, learned augmentation policies such as AutoAugment and RandAugment have emerged and have shown remarkable success in classic computer vision tasks [61, 62]. A key insight from these studies is that the optimal amount of augmentation is dependent on both the model and input dataset size, with smaller models and smaller datasets requiring weaker augmentation policies.

Unfortunately, many of these studies focus solely on 2D image classification problems, ignoring the unique challenges faced in 3D segmentation tasks. For instance, many intensity augmentations that have shown efficacy in RGB imaging cannot be applied to grayscale MR. Moreover, due to data constraints, even the largest datasets for medical image segmentation would be considered extremely small in the computer vision sector. Some studies have looked at augmentation on medical imaging, but are generally not comprehensive. Specifically, these studies tend only to look at the effects of simple transformations utilized in isolation instead of assessing the effect of policies that combine multiple spatial and intensity transformations together [63, 64, 65, 66].

2.2.3 Utilization of Other Datasets to Improve Performance

The goal of supervised machine learning is to learn highly robust and generalizable representations of the input data. When human-annotated labels are scarce, models are unable to learn robust representations, leading to brittle solutions. In the absence of the ability to learn high-quality representations due to dataset size, three main approaches exist: pre-training, self-supervised training, and self-training.

Since many computer vision tasks are similar in nature, it is expected that representations learned on one dataset are transferable to another. Under this assumption, pre-training is the action of training on a large, diverse dataset in order to learn representations which can then be fine-tuned on the dataset of interest [67]. In contrast, self-supervised learning relies only on unlabeled data to learn visual representation. Trivially generated labels are created through a pretext task, and a network is trained on these automatically acquired labels [27]. Pretext tasks can be as simple as predicting the amount of rotation applied to an image, or predicting the relative positioning between two patches from an image. The best approaches leverage data in a task-agnostic way, so as to ensure learned representations are not tailored to any specific task. To that end, current state-of-the-art methods focus on contrastive learning, where the network is simply tasked with distinguishing whether two images are augmented versions of each other [68]. The final paradigm is known as self-training, where a network trained on the task of interest is used to generate pseudo-labels on a separate set of unlabeled data. Following this, a new network is trained from scratch using both the pseudo-labels on the auxiliary dataset and the true labels on the original dataset [69].

While it is well-known that pre-training can produce impressive performance gains in situations where collecting sufficient labeled data is difficult, less is known about how it interacts in the presence of data augmentation [69, 70]. Furthermore, more recent research calls into question the utility of pre-training across domains (i.e. pre-train on classification task, fine-tune on object detection), showing that pre-training across domains confers no advantage (and may in fact be disadvantageous) [67]. Similar issues have been brought up regarding self-supervised learning. Moreover, all state-of-the-art self-supervised methods require millions of unlabeled images in order to effectively learn visual representations, an intractable hurdle in medical imaging[68, 28]. Self-training on the other hand has been shown to be beneficial across dataset sizes and augmentation strengths, and is additive on top of pre-training [69]. To the best of our knowledge, only pre-training has been extensively studied with regards to medical imaging, and the potential benefits stemming from proper application of joint pre-training and self-training has not been explored.

2.2.4 Loss Functions

A loss function is some quantitative measure of the compatibility/similarity between a prediction (what the network outputs) and the ground truth label [19]. During the training process, the weights of the neurons in the network are updated based on the value of the loss function. A large loss implies that the output of the network is very different from what is desired, and as such, the weights of the neurons will change dramatically. Over time, we expect the average loss to go down as the network converges towards the desired result. There are many reasons why a network may not converge (including too high or low of a learning rate for gradient descent, too few parameters, too much regularization, etc.), but often the simplest explanation is that an improper loss function was used. In fact, two dissimilar loss functions can cause otherwise identical networks to have very stark differences in performance.

In general, different tasks require different loss functions (cross-entropy for classification, mean squared error for regression, dice loss for segmentation, etc.). Cross entropy is given by the following equation:

$$H(p,q) = -E_p[log(q)] \tag{2.1}$$

where p is the true distribution (i.e. ground truth) and q is the predicted distribution. Dice score coefficient (DSC) is an intersection over union metric given by the following equation:

$$DSC(p,q) = \frac{2\sum pq}{\sum p + \sum q} \tag{2.2}$$

which measures the degree of overlap between the ground truth shape and predicted shape [71]. DSC ranges from 0 to 1, with 1 representing a perfect overlap. To use as a loss function, we subtract the DSC from 1. For most binary (or multi-class) segmentation tasks, dice loss is the preferred loss function since it exhibits fast convergence and very high accuracy in practice. For example, high precision automatic brain tumor segmentation has been accomplished using dice loss [72].

However, it is important to note that dice loss is an aggregate measure (i.e. for each patient we receive a single loss value), as opposed to a per-pixel loss (such as cross-entropy). This means that as the structure of interest grows larger, small mistakes matter less as they are averaged out across the patient. Overall, this implies that while dice loss handles singular, compact structures such as large tumors very well, it does not provide the accuracy nor the resolution to segment large, intricate structures such as vessel trees. Furthermore, it is not equipped to handle small objects well, since DSC is an unstable metric for small objects. In lieu of using dice loss, performance gains can be realized via the use of specialized loss functions, such as focal loss or boundary-weighted loss, which work by down-weighting easily classified examples [73, 74].

Indeed, the optimization of neural networks in the setting of highly imbalanced data is a challenging task and area of active research. Consider an MRI of a patient with metastatic brain lesions. After applying skull-stripping to remove all but brain tissue, the average size of the volume is about 135 x 165 x 135. Out of these approximately three million voxels, only between 100 and 10000 are tumor. This extreme class imbalance makes training of large models impossible without the use of dedicated sampling heuristics, complex model architectures, or custom loss functions [75, 76, 73].

2.3 Image Segmentation Methods

Image segmentation can be broadly split into two categories: semantic segmentation and instance segmentation [77, 78]. In semantic segmentation, each pixel in the image is given a unique class label. For most medical tasks such as tumor segmentation, this tends to be binary decision between labeling pixels as normal and abnormal. In instance segmentation, each detected object receives its own unique label. In our previous example, this would mean each individual tumor in the image would receive a different class label. An example of semantic vs. instance segmentation for a patient with brain metastases is shown in figure 2-2. Most common methods for segmentation provide only semantic level outputs, and the majority of the work in this thesis does indeed surround semantic segmentation. For this purpose, the following review will focus mainly on such methods, but it is important to note that instance segmentation methods do exist and are becoming more popular with advances in DL.

2.3.1 Classical Approaches

There are a variety of historical approaches to image segmentation and we can loosely categorize them as thresholding based, clustering based, region based, and atlas based. While some of these are now out-dated, many are still actively used and can still provide near state-of-the-art results.

Thresholding Based Approaches

The simplest method for binary segmentation involves splitting the image into two parts at a predefined intensity value, with the goal being to separate the background and foreground classes. One can choose this threshold manually or automatically. Otsu's method is a simple approach to automatically choosing this intensity threshold, whereby if we assume the distribution of background and foreground classes is bimodal,

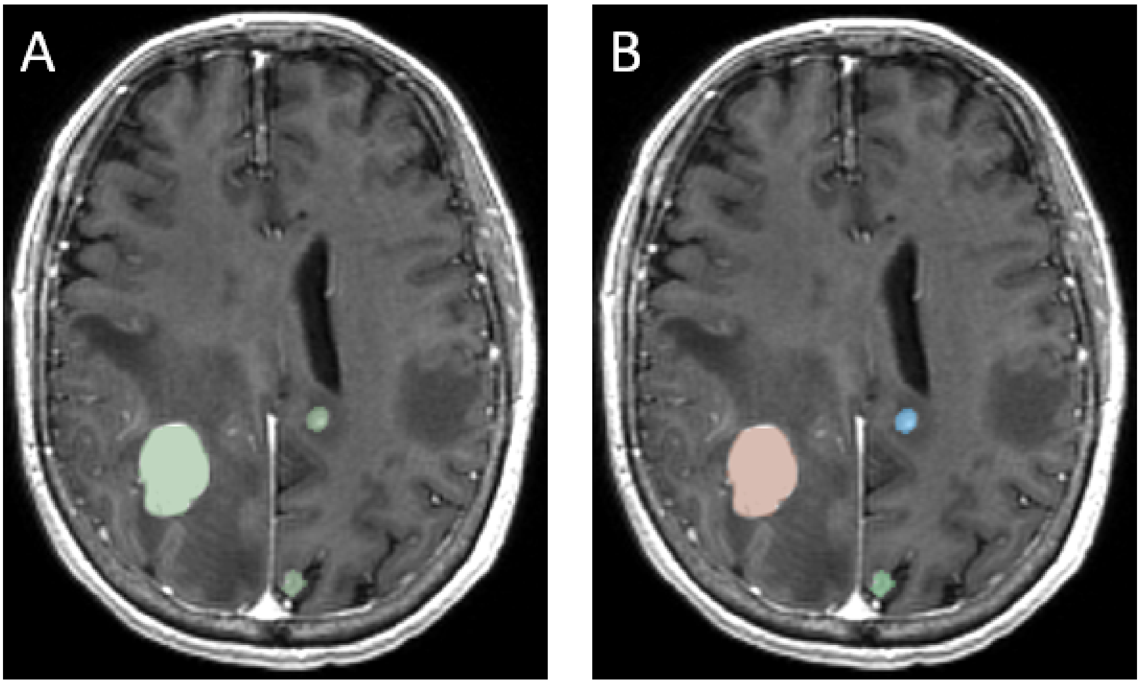


Figure 2-2: **Semantic vs. instance segmentation.** A) We show an example of semantic segmentation of a patient with brain metastases. The three distinct lesions are all given the same label (shown in green in this figure). B) We show an example of instance segmentation for the same patient. Now, each of the three lesions is delineated with a unique label, allowing us to more easily identify the distinct objects (e.g. metastases) for this patient.

we can choose a threshold such that we minimize intra-class intensity variance (or equivalently maximize inter-class intensity variance) [79]. In cases where there is some non-uniform lighting across the image, a global threshold may not work well, necessitating the use of locally adaptive methods [80].

Clustering Based Approaches

Clustering based approaches, such as k-means and Gaussian mixture models (GMM), can be used to split an image into more than two classes [81]. To use a GMM, one begins by randomly initializing n Gaussians, where n refers to the number of desired classes. These randomly initialized Gaussians are then iteratively updated via the expectation maximization (EM) algorithm which finds the maximum a posteriori (MAP) estimates of the Gaussian parameters [82]. For example, this approach can be used to quickly and efficiently segment a brain into three compartments (white matter, gray matter, and CSF). Since clustering based approaches use only the underlying intensity and do not incorporate spatial information, it is often prudent to apply a conditional Markov random fields (cMRF) to spatially smooth the segmentations [83].

Region Based Approaches

Region based approaches utilize both the underlying image intensities as well as taking into account the spatial relationship of nearby pixels. When objects are a known, fairly smooth shape, active contour models (also known as "snakes") are popular [84]. A contour is initialized loosely about the region that needs to be segmented. This contour is then acted upon by internal and external forces that seek to balance how much curvature is allowed in the contour against the intensity gradients in the image. Graph cut based approaches are equally popular, and can be used to segment an image into any predesignated number of classes [85]. By marking down "seed" pixels for each class, the model seeks to find the boundary that maximizes the inter-class intensity gradient. By taking advantage of optimized graph network algorithms, these approaches can be extremely fast in practice [86].

Atlas Based Approaches

If a representative template case exists with all desired classes already segmentated, atlas-based approaches can be used to map the segmentation from the template onto the new image. This mapping can be done in numerous ways, with the most common being via a fully derformable transformation. The Advanced Normalization Tools (ANTs) package [87], has a popular (and highly effective) implementation of atlas based segmentation in the form of the Atropos algorithm [88].

2.3.2 Deep Learning Approaches

There is a wealth of research in DL based image segmentation of natural images [89, 90, 91, 92]. We note that the majority of these methods are designed for 2D images, with relatively less research being done for 3D cases [93]. For the purpose of this background section, we will focus mainly DL based approaches for 3D medical imaging.

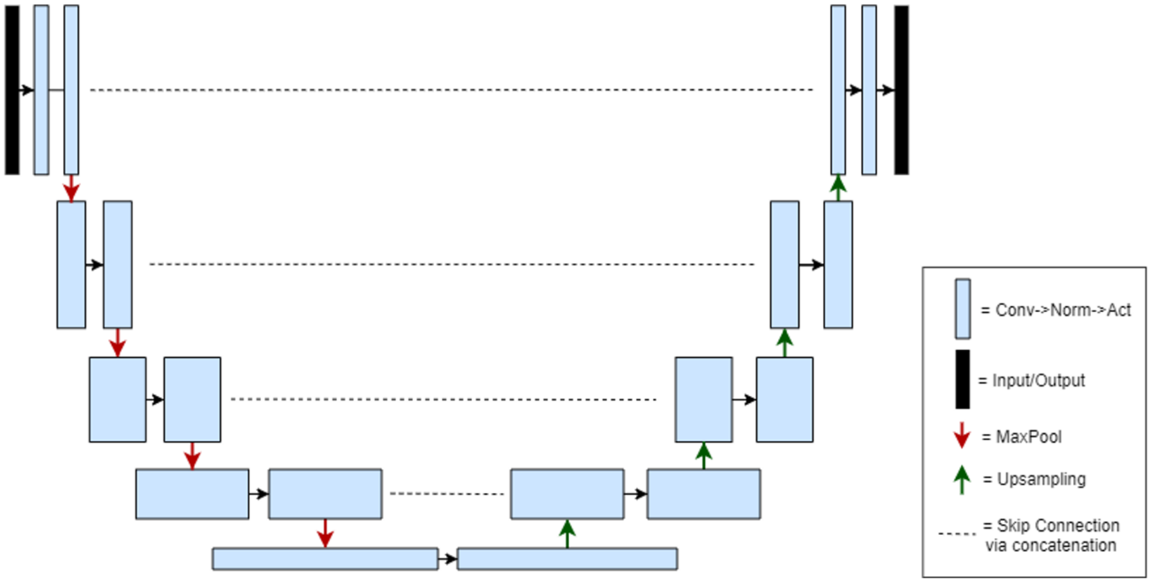


Figure 2-3: **Schematic of a U-Net.** This sample U-Net is composed of 5 levels, with each layer having 2 convolutional blocks. Layers are seperated by downsampling operations (i.e. max pooling, average pooling, strided convolution), or upsampling operations (i.e. trilinear interpolation, deconvolution). Skip connections between the two arms of the network helps propagate gradients and allows the network to shortcut deeper layers as needed.

Perhaps the largest body of segmentation literature as it relates to medical imaging is that for primary brain tumors. This is mainly due to the availability of the large multi-institutional publicly available BraTS dataset [94, 95, 96, 97, 98]. In recent years, 3D U-Net architectures [99] have consistently dominated the BraTS leaderboards and are the current state-of-the-art method for brain tumor segmentation [100, 101, 102].

Briefly, a U-Net is composed of an encoder, which contracts the input image down to a lower-dimensional representation, and a decoder, which expands the lower-dimensional representation back to the original input image size. The encoder is composed of a series of blocks separated by downsampling operations. Each block is composed of one or more convolution operations (along with associated normalization and activation). The decoder is composed in an identical manner, with upsampling operations used in lieu of the downsampling. The difference between a U-Net and a standard encoder-decoder set-up is the addition of skip connections, which allows the network to backpropagate gradients more easily. An example of a U-Net is shown in figure 2-3.

Myronenko won the 2018 BRATS challenge utilizing an asymmetrical residual U-Net, where most of the trainable parameters of the model resided in the encoder. Furthermore, in contrast to the standard U-Net framework which uses four or five downsampling operations in the encoder, he applied only three in order to preserve spatial context [45]. Other modifications to the U-Net structure have also been used with success. Jiang et al. won the 2019 challenge using a two-stage cascaded asymmetrical residual U-Net, where the second stage of their cascade was used to refine the coarse segmentation maps generated by the first stage [103]. The second place that year was awarded to Zhao et al., who utilized dense blocks along with various optimization strategies such as variable patch/batch size training, heuristic sampling, and semi-supervised learning [104]. It is important to note that while architectural modifications to the U-Net can provide performance boosts, they are not always necessary. Indeed, Isensee et al. won the 2020 challenge with their architecture coined ”No New-Net”, highlighting that a vanilla U-Net coupled with excellent training and optimization strategies can still achieve state-of-the-art results. Moreover, they achieved an average testing set dice score of 88.95% for whole tumor segmentation, achieving segmentation performance indistinguishable from human experts [101]. in 2021, a team from Nvidia took this one step further by running a large set of ablation studies to figure out which components of the U-Net architecture were most important for segmentation. Notably, while most networks use 5 or at most 6 levels, they find that going deeper to 7 levels improved performance [102]. Nvidia also took the second spot in 2021, with a similar U-Net based approach [105].

With such strong competition results, it is thus unsurprising that U-Nets are used for most medical image segmentation projects [106]. For these reasons, all of the segmentation work in this thesis use 3D U-Nets as well.

2.4 Image Registration Methods

There is extensive literature in image registration and it can be broadly be split into two categories: non-learning based and learning-based. We begin by explaining some terminology and then provide a brief overview of relevant methodology.

2.4.1 Registration Problem

Given a fixed and a moving image, the goal of image registration is to find a mapping such that the moving image is transformed into the fixed image [107]. This mapping can be parameterized as either a linear or non-linear transformation, and can be done in either 2D or 3D. Registration can be uni-modal (MRI to MRI), multi-sequence (T1-post to FLAIR), or multi-modal (PET to CT).

2.4.2 Linear Transformations

A linear transformation (also known as an affine transformation) is composed of a set of rotations, translations, scales, and shears [108]. For 2D registration, there are 7 parameters total: one rotation parameter rot, 2 translation parameters $trans_x$ and $trans_y$, 2 scale parameters $scale_x$ and $scale_y$, and 2 shear parameters $shear_x$ and $shear_y$. These parameters can be combined together into a single transformation as follows:

$$\begin{aligned}
Affine_{2D} = &\begin{bmatrix} \cos(rot) & -\sin(rot) & 0 \\ \sin(rot) & \cos(rot) & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} scale_x & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
&* \begin{bmatrix} 1 & 0 & 0 \\ 0 & scale_y & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & shear_x & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 0 & 0 \\ shear_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
&* \begin{bmatrix} 1 & 0 & trans_x \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & trans_y \\ 0 & 0 & 1 \end{bmatrix}
\end{aligned} \tag{2.3}$$

We note that because matrix multiplication is not commutative, the order which we apply the individual transformations matters. For example, a rotation followed by a translation is not the same as a translation followed by a rotation. For 3D registration, there are 15 parameters total: three rotation parameter rot_x, rot_y, and rot_z, 3 translation parameters $trans_x$, $trans_y$, and $trans_z$, 3 scale parameters $scale_x$, $scale_y$, and $scale_z$, and 6 shear parameters $shear_{xy}$, $shear_{xz}$, $shear_{yx}$, $shear_{yz}$, $shear_{zx}$, and $shear_{zy}$. Combining the individual transforms into a 3D affine matrix follows similarly to the 2D case.

In order to apply an affine transformation matrix to an image, it is first converted into a dense displacement vector field (DVF). Letting n be the number of dimensions in your image, at each coordinate in the vector field, a vector of length n defines how far the pixel at that point will move. An interpolation algorithm is used to interpolate the resultant grid points, giving the newly transformed image. An example of various affine transformations (along with the associated DVFs) is shown in figure 2-4. Grid lines are overlaid on each image in the figure. Since affine transforms are purely linear, all gridlines remain straight and parallel with each other. We note that translation, rotation, and shearing are area preserving transforms. More specifically, the determinant of the matrix representing those transforms is 1, indicating no volume change. This is of course not the case for scale transforms, which are the only affine component that can change a structures volume. This can be visualized in the figure by noting that that the area of each gridline box in the scale transformed image is larger than it is in the other three.

Linear registration is a necessary part of the medical imaging pipeline [109, 110]. The most common use cases are to co-register images acquired at different patient timepoints (e.g. baseline to follow-up) or different viewpoints (e.g. the same patient images at slightly different positions or using different modalities, scanners, or protocols). For instance, for brain tumor imaging, it is routine to acquire T1-post contrast, T1-pre contrast, T2, and FLAIR imaging sequences, as they provide different, but complementary information about the tumor.

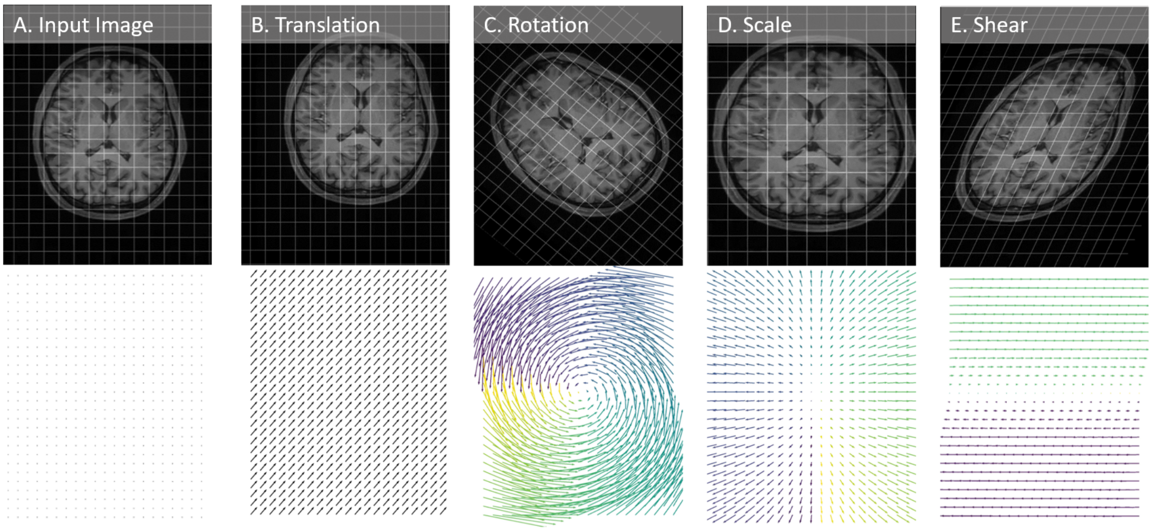


Figure 2-4: **2D affine transformations with displacement vector fields.** A) A sample brain image is shown with an identity DVF. Grid lines are overlaid on top to visualize how the transform affects the underlying grid. B) An example of translation in the positive x and y directions. C) An example of counterclockwise rotation about the origin. D) An example of scaling up the image anisotropically. E) An example of shearing along the x axis.

2.4.3 Non-linear Transformations

While a linear transformation can be represented as an affine matrix (or a DVF), a non-linear transformation (also known as a deformable transformation) can only be represented as a DVF [111]. This is because deformable transforms can move pixels independently of each other, meaning there is no global transform being applied. While deformable transformations can represent any type of one-to-one mapping, they are often constrained in certain ways to ensure that the output image remains recognizable after being deformed. Spatial smoothness constraints on the DVF can prevent physiologically impossible transformations such as folding or creasing. To quantify this, the jacobian, which is a vector of the n first-order partial derivatives of the DVF, can be calculated, with the determinant of the jacobian at each point defining the local volume change [112]. A negative jacobian determinant at a specific point indicates that the transformation locally at this point resulted in a fold. An example of a deformable transformation (along with the associated DVF) is shown in figure 2-5. Looking at the deformably moved image, it is clear to see that the gridlines are no longer straight and parallel to each other, with some areas experiencing volume growth and others volume shrinkage. Because of our DVF is spatially smooth, the jacobian determinant is non-negative everywhere.

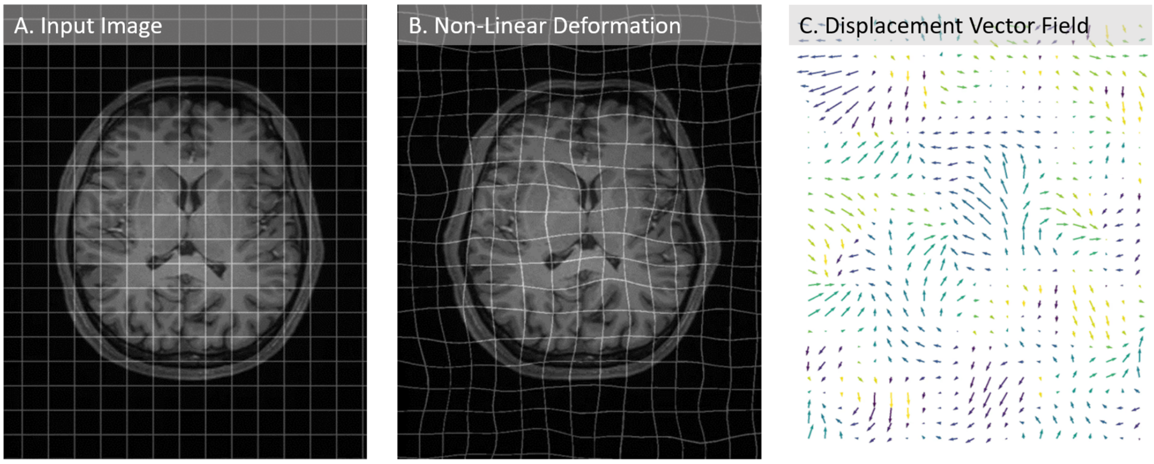


Figure 2-5: **2D deformable transformation with displacement vector field.** A) A sample brain image is shown. Grid lines are overlaid on top to visualize how the transform affects the underlying grid. B) An example of a deformably transformed image. The gridlines are no longer straight and parallel to each other, showing how different parts of the image are experiencing different local volume changes. C) The associated DVF of the deformable transformation.

Non-linear registration has many use-cases in medical imaging, especially when large deviations are expected between images [113]. For instance, aligning images from different patients in order to create a template atlas requires deformable methods. Even for intra-patient registration, large deviations can be seen due to disease progression, respiratory effects, and/or weight gain/loss, among other things.

2.4.4 Sequential Image Registration

Sequential refers to solving lower complexity transforms before higher complexity transforms. In other words, a purely affine transformation is computed first before solving for the deformable transformation. While the deformable registration can in theory represent both linear and non-linear transformations simultaneously, the optimization problem becomes significantly more difficult and may even become intractable. An example of sequential image registration is shown in figure 2-6. In this example, a fixed and moving image are separated by some unknown linear and non-linear transformation. First, we find the purely linear transformation separating these two cases. As we can see in the figure, some combination of rotation and translation (with a small bit of anisotropic scaling) is enough to approximately match the two images. Using this affinely aligned image as initialization, we subsequently run deformable registration. The final output image exactly matches the fixed image, indicating a high quality registration. While perfect deformable registration may be possible with synthetically shifted images, it is less likely on real-world data. This is especially true when there are large differences between images, such as those stemming from anatomical abnormalities (i.e. tumors, diseases, surgical insertions/removals, etc.).

2.4.5 Non-Learning Based Registration

Given a fixed and a moving image, classical registration approaches perform a gradient descent based numerical optimization to iteratively align pixels from the moving image onto the fixed image to improve a chosen similarity metric (e.g. mean squared error (MSE), normalized cross correlation (NCC), etc.). The learned transformation can be either linear or non-linear, depending on one's use case. To alleviate this numerical optimization problem (which even for linear transforms can get stuck at poor local minimas for anatomically complex images), classical methods often employ a *sequential* and *pyramidal* hierarchy. *Pyramidal* refers to a multi-scale approach wherein the transformation is first computed at a coarser image scale and is progressively updated at finer image scales [114]. We note that due to the iterative nature of these classical algorithms, they can be quite computationally intensive. Indeed, deformable registration of 3D brain MR imaging can take upwards of one to two hours on CPU per image pair. There are many classical registration algorithms for deformable registration, including but not limited to B-splines [115], Demons [116], Large Diffeomorphic Distance Metric Mapping (LDDMM) [117], and Symmetric Normalization (SyN) [118]. The SyN algorithm from the ANTs package [119] is the current gold standard as it is generally regarded as the best classical deformable

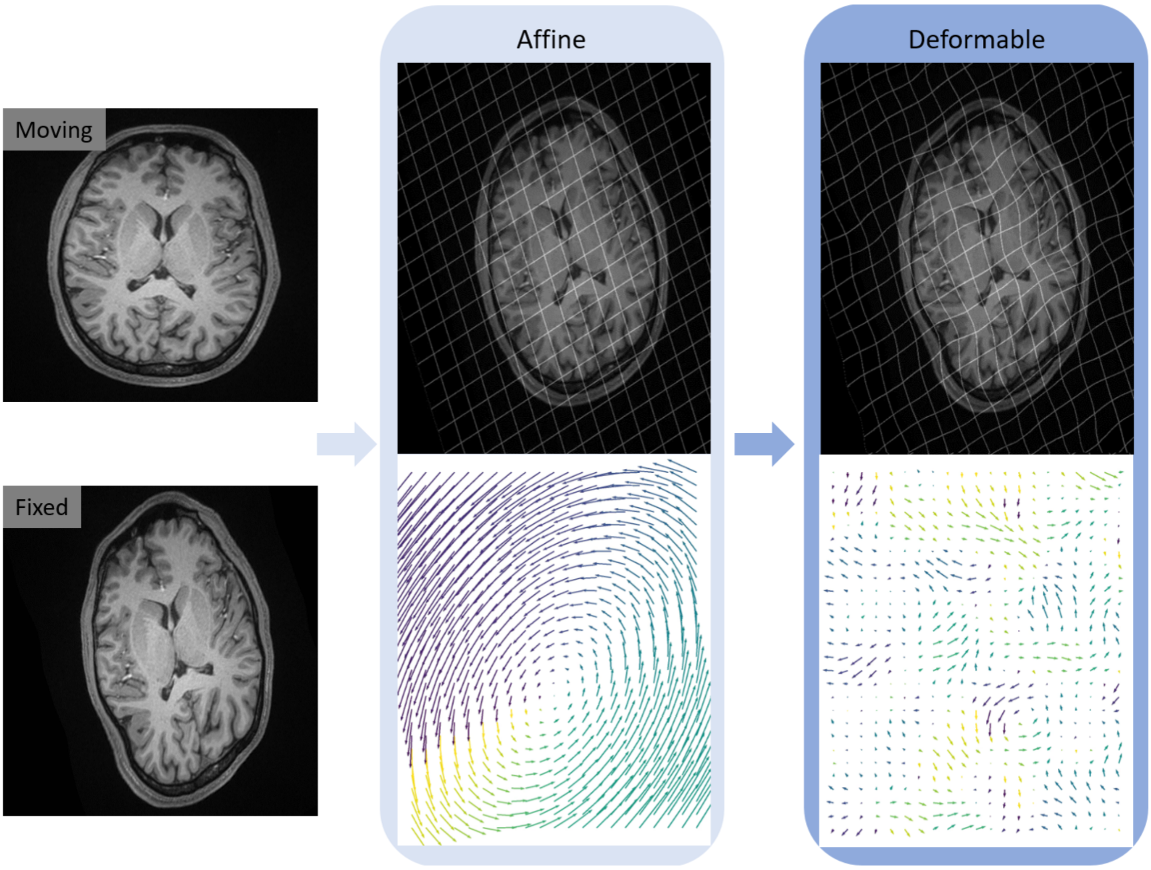


Figure 2-6: **Affine and deformable image registration example.** Our goal is to register the moving image to the fixed image. We solve this sequentially, starting with a purely affine registration. We then use the affinely moved image as initialization for the deformable registration algorithm, which computes the final registration and associated DVF.

registration algorithm [120].

2.4.6 Learning Based Registration

Newer methods utilize neural networks to learn a function for (affine and/or deformable) registration. This can be advantageous because each image pair can be fully registered with one forward pass of the network, which will take only a few seconds on GPU. DL based affine registration networks are usually formulated as a supervised regression problem. Chee et al. use a *Siamese* style encoder to directly predict the affine transform matrix [121]. Islam et al. uses a similar regression approach, but focuses on cross-modality registration [122]. DL based deformable registration networks can either be trained in a supervised or unsupervised manner. While earlier approaches like that from Sokooti et al. required ground truth DVFs to train the network [123], newer approaches tend to be fully unsupervised. Dalca et al. proposed a U-Net based diffeomorphic registration model they named VoxelMorph (VXM) [124]. Building off this approach, Mok et al. utilized a pyramidal architecture to improve the quality of the registration [125]. However, they did not incorporate feature sharing at the different levels of the pyramid, resulting in redundant parameters. de Vos et al. devised a network to sequentially perform affine and deformable registration, but their deformable registration was based only on b-spline grids [126]. Christodoulidis et al. also performed both affine and deformable registration, but their approach was neither sequential nor pyramidal [127].

2.5 Adrenal Gland Segmentation

Incidental adrenal masses are common in adults, with an estimated prevalence ranging from 3% to 7% [128, 129, 130]. Abnormalities of the adrenal gland can generally be categorized as either hyperfunctioning or nonfunctioning masses [130]. Hyperfunctioning lesions produce an excess of hormones (cortisol, aldosterone, epinephrine, etc. depending on the specific abnormality), causing chemical imbalance in the body. Conversely, nonfunctioning masses produce no significant change in hormone levels, but cause visible enlargement of the gland. It is important to note that the lack of function does not immediately rule out malignancy. For patients with some form of adrenal abnormality, an estimated 80% are benign and non-functioning, thus never affecting the patient over the course of their lifetime [131]. Insufficient attempt to separate clinically relevant from insignificant disease leads to over-diagnosis, which can lead to increased stress and anxiety, potential harm stemming from unnecessary diagnostic procedures, and substantial financial burden to the patients [132]. Thus, the primary goal of management of incidentally found adrenal masses is to correctly identify the masses that are either 1) malignant or 2) hyperfunctioning in order to spare the majority of patients from requiring further treatment [132].

Identification of potentially suspicious lesions can be done by assessing various radiographic and clinical features in tandem. Briefly, lesions that are large (>4cm), growing (compared to previous time points), have irregular boundaries, and exhibit contrast enhancement are more likely to be malignant. Clinical information including previous/current history of cancer can inform decision making as well. Qualitative assessment of features such as boundary irregularity is highly subjective, leading to inconsistent decision making. Literature regarding robust automated quantitative assessment of adrenal masses is sparse, and to the best of our knowledge, only one study has looked at automated adrenal mass segmentation (and it utilized MR instead of the much more commonly acquired CT) [133].

2.6 Brain Metastases Segmentation

Brain metastases are among the most common intracranial lesion in adults, with an estimated 170,000 new cases diagnosed in the US every year [134, 135]. Historically, overall survival for patients was low, but recent advances in systemic and targeted therapeutics has improved prognosis and prolonged time to neurologic dysfunction [134]. In this context, effective analysis of longitudinal MR is one of the backbones for assessing tumor response in patients with brain metastases [136]. In current clinical practice, patients receive surgery, radiation, and/or chemotherapy, and undergo MR scans at regular intervals throughout their therapy. To assess efficacy of the current treatment regimen, neuroradiologists track individual lesion sizes across time points [137]. If non-negligible enlargement of metastatic lesions is noticed overtime, a different treatment option may be needed. However, as many patients can have more than 10 lesions, manual delineation of all metastatic lesions is prohibitively time-consuming. As a result, most radiographic response criteria, including the RANO-BM, restrict response assessment to a select number of target lesions [136]. This fails to capture the full extent of the disease burden and can miss changes over time in non-target tumors. In addition, manual segmentation is subject to large amounts of inter-rater variability and metastases can range from less than 0.1 mL (i.e. micro-metastases) to greater than 10.0 mL while having varied shapes/structures (from spherical to highly irregular), making consistent outlining challenging [138]. To better capture the full intracranial disease burden, automatic segmentation approaches can be used. Due to the aforementioned variability in size and shape of lesions, current automatic methods suffer from either poor detection of micro-metastatic lesions or high false positive rates (in order to capture micro-metastatic lesions) [139, 140].

Chapter 3

Adrenal Gland Segmentation and Classification

Incidental adrenal masses are common in adults, with an estimated prevalence of about 6%. Despite standardized reporting systems and strict guidelines for the definition of an adrenal mass, there exists significant inter-rater variability for this task. While over-diagnosis can lead to unnecessary clinical workups and follow-up examinations, under-diagnosis risks missed detections of clinically significant disease. In this chapter, we discuss a novel deep learning algorithm that segments adrenal glands on contrast-enhanced CT images and classifies them as either normal or mass-containing. We assess both automated segmentation and classification performance, and present an exploratory analysis on the amount of inter-rater variability present in clinical practice.

3.1 Introduction

Adrenal masses are common, occurring in 6% of the population in an autopsy series and approximately 4% of all abdominal CT examinations [141, 142]. In patients without a history of cancer, the vast majority of adrenal masses that are detected incidentally at cross-sectional examinations are benign [143, 144]. However, some masses may require additional characterization with use of imaging (adrenal mass protocol CT, MRI, PET/CT, or follow-up examinations), biochemical evaluation, biopsy, or surgical excision depending on the size, imaging features, and hormonal activity [132, 145, 146]. When an adrenal mass is detected, subsequent recommendations on whether additional testing is needed is often determined by whether the mass is new, enlarging, or stable compared with prior imaging. Given that management is dependent on accurate interpretation of prior imaging, decreasing radiologist variability for adrenal mass detection may reduce ambiguity and unnecessary follow-up.

Machine learning has been proposed as a strategy to help automate image analysis and improve diagnostic performance [147]. Supervised machine learning algorithms require accurate annotation by labeling and/or contouring of the anatomic structure of interest. To reduce the substantial interreader variation in manual contouring of anatomic structures, automated segmentation approaches are becoming more commonplace for medical tasks. There have been recent machine learning methods to automate segmentation of abdominal organs at both CT and MRI [148, 149, 150, 151, 152, 153]. However, there has been less research performed specifically on adrenal segmentation methods [154, 155]. Additionally, only a few studies have used machine learning approaches for adrenal mass characterization [156, 157]. Therefore, there is a need for greater standardization regarding adrenal gland segmentation and classification.

The purpose of this study was to create a machine learning algorithm that accurately segments and differentiates normal glands from those containing masses at contrast-enhanced CT and to assess algorithm performance. Examples of adrenal glands are shown in figure 3-1. As can be seen, glands can present with many different shapes and sizes, making the differentiation of normal from mass a difficult task.

3.2 Materials and Methods

This retrospective study was compliant with the Health Insurance Portability and Accountability Act and was approved by our institutional review board; the need for informed consent was waived.

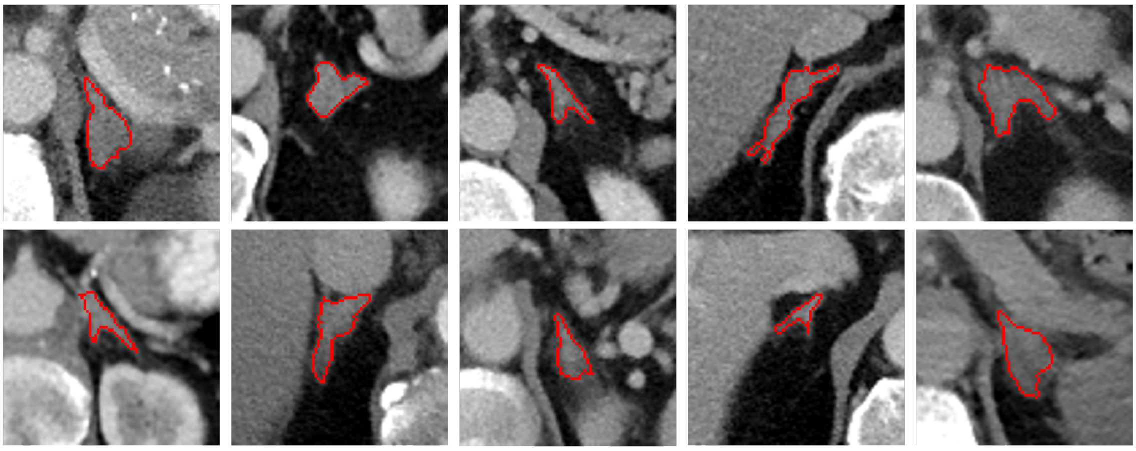


Figure 3-1: **Examples of adrenal glands on CT imaging.** Adrenal glands can take on many different shapes and sizes, with some examples shown here.

3.2.1 Development, Secondary, and Tertiary Test Data Sets

This study included two groups of patients who underwent portal venous phase abdominal CT at Massachusetts General Brigham (MGB), a large academic health system that performs nearly 2 million radiology examinations annually. The first group (hereafter referred to as the development data set) comprised consecutive portal venous phase contrast-enhanced CT examinations from January 1 to January 5, 2012 (figure 3-2). Patients with an adrenal mass were excluded. The development data set was intentionally enriched with additional CT examinations that depicted adrenal masses to increase the proportion of adrenal masses in the training data set. These additional examinations were identified with use of a natural language processing tool (NLP) to search for the word "adrenal" in the Impression section of contrast-enhanced abdominal CT reports from January 1 to December 31, 2012. These reports were manually reviewed and the corresponding images were verified to confirm the presence of an adrenal mass larger than 10 mm in the short axis. The study sample selection strategy was based on prior experience training segmentation neural networks. Specifically, a training sample of at least 200 examinations was desired to ensure a robust and generalizable network, with an additional 25 examinations apiece for validation and testing. In addition, a large number of examinations with adrenal masses were required to ensure good performance of the model. Thus, the development data set included 170 normal examinations and 104 examinations depicting an adrenal mass.

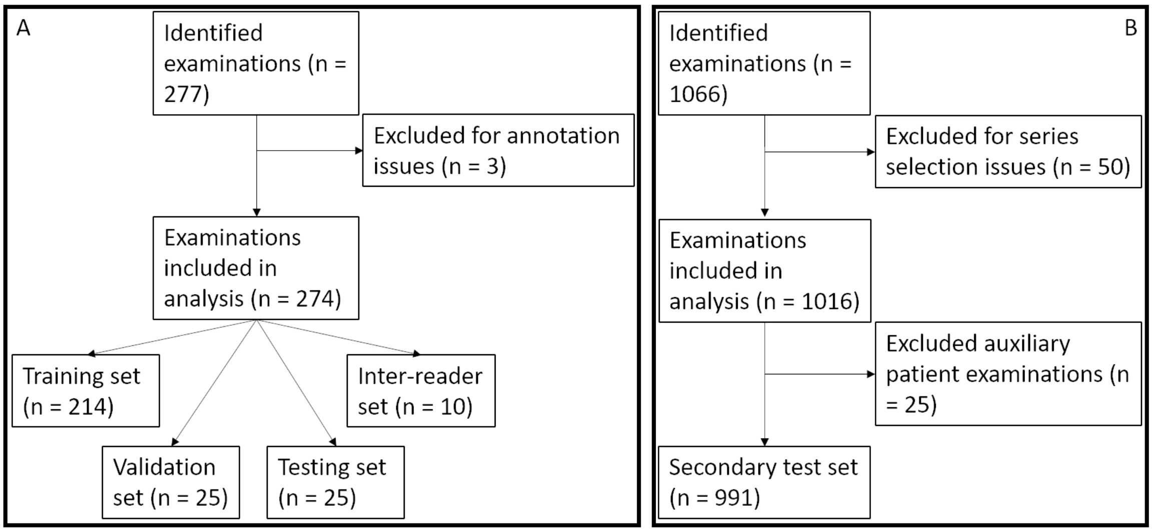

A
Identified examinations (n = 277)
Excluded for annotation issues (n = 3)
Examinations included in analysis (n = 274)
Training set (n = 214)
Inter-reader set (n = 10)
Validation set (n = 25)
Testing set (n = 25)
B
Identified examinations (n = 1066)
Excluded for series selection issues (n = 50)
Examinations included in analysis (n = 1016)
Excluded auxiliary patient examinations (n = 25)
Secondary test set (n = 991)

Figure 3-2: **Inclusion and exclusion criteria flowchart.** A) Development data set flowchart. Three images were excluded due to annotation issues. B) Secondary test set flowchart. Fifty images were excluded due to series selection issues, and 25 images were excluded due to being follow-up imaging.

The second group (hereafter referred to as the secondary test set) consisted of consecutive portal venous phase contrast-enhanced CT examinations performed from November 1 to December 31, 2019. Duplicate patients were excluded from this group. Basic demographic data were collected from the electronic health record (Hyperspace, Epic). Including both the development and secondary test group, 251 of 1242 patients have been previously reported in a prior study [158].

A final group (hereafter referred to as the tertiary test set) consisted of consecutively acquired imaging with no exclusions. This dataset was mainly used to assess interreader classification variability during routine clinical practice.

3.2.2 CT Acquisition

Examinations were performed on a variety of different multi–detector row CT scanners (Siemens Somatom Definition AS, Somatom Definition AS+, Somatom Force, and Somatom Perspective and Toshiba Aquilion One) during the study period. All examinations used a fixed delay of 70 seconds following intravenous contrast material administration, and the series selected were typically acquired with a kilovoltage peak of 120 at inspiration. Tube current modulation was used, and axial section reconstruction was 5 mm. The intravenous contrast agent was iohexol (Omnipaque, GE Healthcare). Patients weighing less than 150 pounds (68 kg) received 75 mL, and patients weighing 150 pounds or more received 100 mL.

3.2.3 Image Annotation

All CT examinations were viewed and annotated in a commercially available picture archiving and communications system (Visage Imaging, version 7.1.15). Radiologists were blinded to clinical history and patient information. For 264 examinations from the development set, adrenal gland segmentation was performed by one of five radiologists (one resident [B.D.], one abdominal imaging fellow [C.R.W.], and three fellowship-trained radiologists [D.I.G., B.C.B., and W.W.M-S., with 6, 12, and 28 years of experience, respectively]). On the remaining 10 examinations, all five radiologists performed adrenal segmentation of 19 glands (one surgically absent) to measure interreader variability.

3.2.4 Adrenal Classification

Masses were defined by the radiologists as space-occupying lesions not conforming to the normal shape of the adrenal gland and measuring 10 mm or greater in the short axis, an example of which is shown in figure 3-3 [132]. If the patient had undergone an adrenalectomy, the side was marked as “resected” and was not segmented. Adrenal classification (normal or mass) for the development set was provided by one of five radiologists. Classification for the secondary set was provided by one of two board-certified, fellowship-trained abdominal radiologists (D.I.G. or W.W.M-S.) to serve as a reference standard. The largest-diameter circle that would fit completely within the ground truth segmentation was calculated for the development set of 274 examinations. This automatic diameter measurement was compared with the radiologist classification, noting the percentage of radiologist-classified normal adrenal glands with diameter measurements 10 mm or greater and masses less than 10 mm.

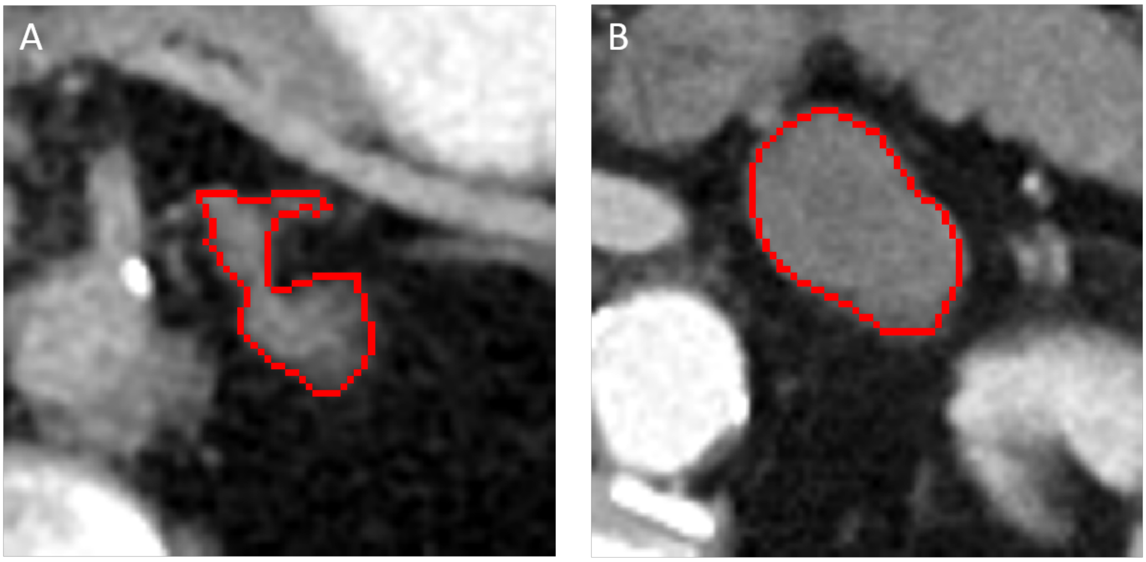


Figure 3-3: **Examples of adrenal masses.** A) Adrenal masses must not conform to the normal shape of the adrenal gland. In this example, a nodule is visible on the lateral limb of the gland. B) Adrenal masses should measure 10 mm or greater in short axis. In this example, a large mass is visible.

3.2.5 Machine Learning Algorithm

Our adrenal mass detection and classification pipeline included two stages. First, a three-dimensional segmentation convolutional neural network U-Net [99] was used to identify the pixels in the full CT examination volume that represented the adrenal glands. Second, a cropped area around each adrenal gland was passed to a three-dimensional classification convolutional neural network, DenseNet [159], which predicted whether or not the input image contained an adrenal mass. An overview of the full process is shown in Figure 3-4. The source code for this study can be found at *https://github.com/QTIM-Lab/AdrenalMGB-Version-1*. The code identifier is AdrenalMGB-Version-1. Full details about the machine learning algorithm follow in the subsequent sections.

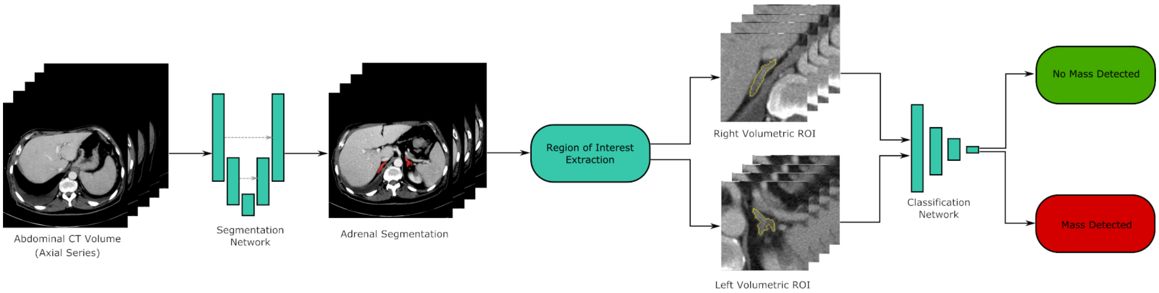


Figure 3-4: **Overview of the adrenal segmentation and classification process.** An axial CT series is preprocessed and passed to a U-Net for segmentation of the adrenal glands. Regions of interest (ROIs) are cropped around the adrenal glands, extracted, and passed to a classification network to determine whether the gland contains a mass.

Image Preprocessing

To mitigate variability between patients stemming from differences in imaging protocols, we apply the following preprocessing steps. First, to reduce the field of view, space was cropped from around the body in all images by thresholding the CT image at -500 Hounsfield Units (HU), followed by a series of morphologic hole-filling operations to extract a smooth contour of the body. The body cropped series was then resampled to a fixed voxel spacing of $1 \times 1 \times 5$mm, and finally windowed using the level/width setting of $70/370$ HU, a setting that visualizes soft tissue well. Window level setting is an important step in CT interpretation and is used by radiologists to enhance the visualization of certain pathologies [160, 161]. In figure 3-5, we show an example CT with and without soft-tissue windowing. It is significantly harder to discern boundaries between different soft-tissue structures in this image than it is in the windowed image. The left adrenal gland in particular is difficult to visualize when no windowing is applied.

Segmentation CNN Architecture

We used a five-level 3D U-Net [99] which takes the windowed CT image as input and outputs a binary map delineating the adrenal glands. Each level in the down-sampling and up-sampling paths of the network is composed of two convolution blocks defined as follows: $3 \times 3 \times 3$ convolution -> group normalization operation [162] -> nonlinear

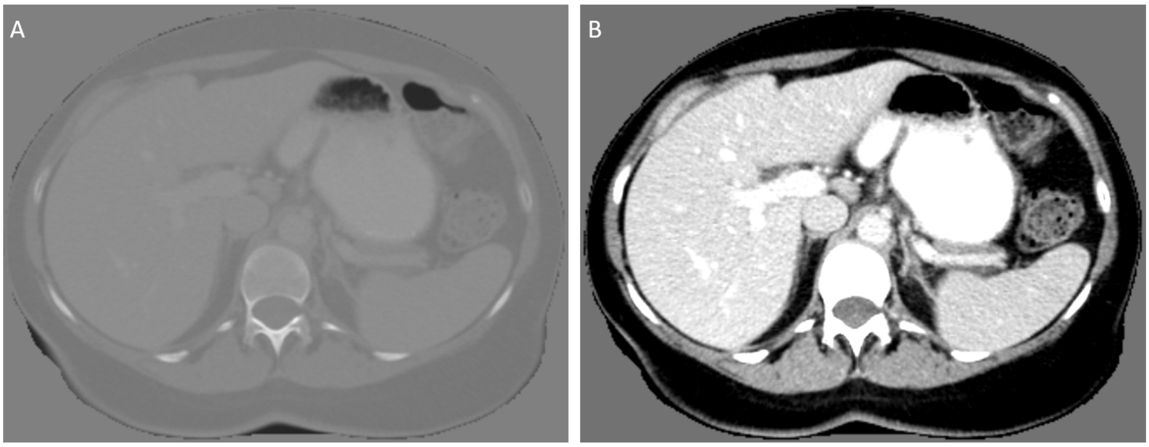


Figure 3-5: **Example of CT image with and without windowing.** (A) A non-windowed CT image. While it is possible to make out the major anatomical organs such as the liver, the difference in contrast between soft-tissue structures is low. (B) After applying a level/width setting of 70/370 Hounsfield Units (HU), we can discern anatomical boundaries much better. Indeed, the adrenal glands in particular are easier to identify.

activation function (ReLU) [163]. Feature map down-sampling and up-sampling is accomplished through strided convolution (with stride of 2) and trilinear interpolation, respectively. To ensure our model learns sufficient anatomic context to correctly localize the adrenal glands, we train our network on patches of size $224 \times 224 \times 32$ voxels (the median size of the preprocessed CT images is $392.5 \times 283 \times 90$ voxels). Our network uses 32 filters in the convolutions in the first level, and we double the number of filters in each level as we go deeper into the network. Group Normalization (with group size of 16) [162] was used in lieu of Batch Normalization (BN) [164] to alleviate the effects of the small batch size necessary to train a 3D model on large patches. To encourage faster convergence and ensure that deeper layers of the decoder are learning semantically useful features, we employ deep supervision by integrating segmentation outputs from different levels of the network [165, 104].

Segmentation CNN Optimization

Due to the significant class imbalance present in the dataset (the ratio of segmented adrenal gland voxels to background is $\approx 0.02\%$), we use biased sampling procedures during training. Specifically, we sample patches from the series such that 50% of patches contain segmented adrenals, and force all batches to contain at least one adrenal patch. Training was performed with a batch size of 2 using the stochastic gradient descent optimizer with decoupled weight decay and momentum set to 0.9 [56]. We progressively decrease the learning rate and weight decay using a cosine anneal schedule with an initial learning rate and weight decay of 0.2 and 0.00002, respectively. These values are decreased by a factor of 200 over the course of approximately 500 epochs. The loss function is the unweighted sum of the DSC loss and cross-entropy loss. To mitigate overfitting, we apply real-time data augmentation during the training process. Specifically, spatial transformations included random anisotropic scaling (0.75 to 1.35), rotations ($-20°$ to $20°$), shearing (-0.20 to 0.20), and translations (-20 to 20 pixels) around all three axes, as well as random elastic deformations. Intensity augmentation in the form of gamma correction (.85 to 1.15) was also used, after HU intensity windowing. All augmentations were applied with probability of 0.5, with the exception of elastic deformations, which were applied with probability of 0.1.

Segmentation CNN Inference

We trained a total of 5 segmentation networks as detailed above and used these 5 networks as an ensemble. Since the CT volumes are significantly larger than our training patch size, we adopt a sliding window approach, where each tile is equal to our training patch size ($224 \times 224 \times 32$) and adjacent tiles overlap by a certain amount. Greater patch overlap improves segmentation quality at the cost of greater computational burden. Since the adrenals cover only a very small portion of the CT volume, we first run a coarse segmentation with a patch overlap of 10% to localize the adrenal glands. A high patch overlap of 85% in that small region of interest is then used to refine the predicted segmentation. On average, inference for the ensemble using this approach takes 3 minutes per patient.

Segmentation CNN Postprocessing

Simple postprocessing was applied to ensure that only the two largest connected components remain, which were then split into separate right and left adrenal masks.

These masks were subsequently centered and cropped to an 80 × 80 × 24 bounding box and then passed into the classification CNN detailed below.

Classification CNN Architecture

We used a modified 3D DenseNet architecture [159] which takes a windowed CT image and the predicted segmentation mask from the previous network of size 64 × 64 × 16 as input and outputs a probability that the adrenal gland contains a mass. A layer in the network was defined as follows: 1 × 1 × 1 convolution -> BN [164] -> ReLU [163] -> 3 × 3 × 3 convolution -> BN -> ReLU. The full network is comprised of 4 dense bottleneck blocks, with 8, 8, 16, and 32 layers in the first, second, third, and last block, respectively. The growth rate k of all layers in the network was 32 and the bottleneck convolution reduces the number of input feature maps to $4k$ maps. Transition layers between the blocks used max pooling [166].

Classification CNN Optimization

A class imbalance was also present for this task, with the ratio of masses to normal glands being approximately 22:100. To mitigate this effect, we oversample from the positive class each epoch, reducing the class imbalance to approximately 38:100. Training was performed with a batch size of 16 using the stochastic gradient descent optimizer with decoupled weight decay [56] and momentum set to 0.9. We progressively decreased the learning rate and weight decay using a cosine anneal schedule with an initial learning rate and weight decay of 0.1 and 0.0005, respectively. These values were decreased by a factor of 250 over the course of approximately 250 epochs. The loss function was the weighted binary cross-entropy loss, where the relative weight of the positive class to the negative class was 4. The same data augmentations used to train the segmentation network were used here.

Classification CNN Inference

We trained a total of five classification networks as detailed above and use these five networks as an ensemble. Test time augmentation is applied by generating 32 distinct patches of size 64 × 64 × 16 from the 80 × 80 × 24 input images, and averaging across the network outputs. On average, inference for the ensemble using this approach takes about 4 seconds per adrenal gland for each patient.

3.2.6 Image Mosaics

To more quickly visualize model outputs, we create mosaic images. Specifically, after running both the segmentation and classification models, we take the centered and cropped bounding boxes of size 80 × 80 × 24 and plot each z-slice of the image into a rectangular montage image. We also draw an automatically computed maximum diameter measurement onto the mosaic image. To calculate this maximal diameter measurement, we start by taking the euclidean distance transform (EDT) [167] of the 2D segmentation on each z-slice of the mosaic, given by:

$$EDT(x) = \min_{y \in Y} \|x - y\|_2^2 \tag{3.1}$$

where Y is the set of all points on the boundary of the segmentation mask, and $x \in X$, where X is the set of all points inside the boundary of the segmentation mask. The maximal value of $EDT(x)$ indicates the radius of the largest circle that can be inscribed into the shape. We find the largest such circle across all 24 z-slices of the mosaic, and draw that onto the mosaic. These diameter measurements are important since size is correlated with abnormality. Glands with maximal diameters greater than 10mm are more likely to be a mass than glands with diameters less than 10mm. Example mosaics are shown in figure 3-6. Panel (A) shows a right sided normal adrenal gland (with maximal diameter measurement of 8.9mm) and panel (B) shows a left sided adrenal mass (with maximal diameter measurement of 15.6mm).

3.2.7 Statistical Analysis

The primary outcomes were (a) agreement of the algorithm segmentation of the adrenal glands compared with radiologist-generated contours and (b) ability of the algorithm to classify adrenal glands as normal or mass-containing. Model segmentation performance

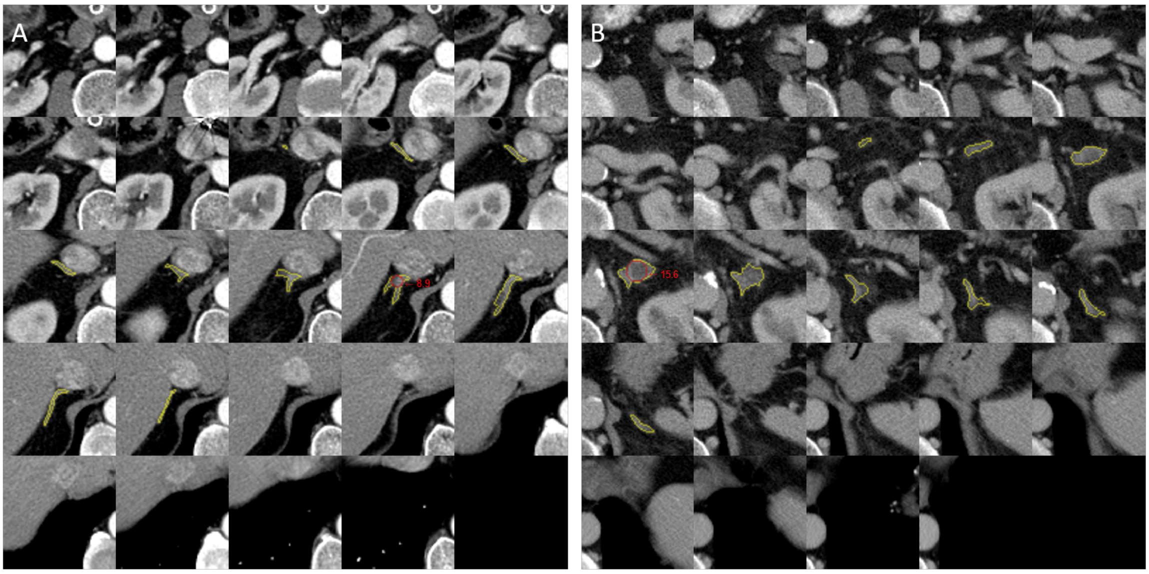


Figure 3-6: **Example mosaic images of a normal gland and an adrenal mass.** (A) Right sided normal adrenal gland. (B) Left sided adrenal mass of approximately 15.6mm. Adrenal glands are contoured in yellow. Largest diameter measurement (which if greater than 10mm is indicative of an adrenal mass) is contoured in red.

was evaluated with use of the Dice similarity coefficient (DSC) [71]. A two-sample t test was used to compare interreader DSC with model DSC. A DSC of 1.0 represents complete overlap between the ground truth and model segmentations, and a DSC of 0.0 represents no overlap. Model classification performance was evaluated with use of sensitivity and specificity, with radiologist categorization as the reference standard. Cohen κ was used to measure agreement between the radiologist and model while accounting for the possibility of agreement occurring by chance [168]. All statistical analyses were performed using Python 3.6.9 [169]. Statistically significant difference was set at $P \leq .05$.

3.3 Results

3.3.1 Patient and Data Set Characteristics

Patient demographics and data set characteristics are summarized in Table 3.1. The development set consisted of 170 consecutive contrast-enhanced abdominal CT examinations without adrenal masses performed at MGB. This group was then enriched with an additional 107 examinations that depicted adrenal masses identified using NLP, excluding cases with adrenalectomy, nephrectomy, significant anatomic or post-surgical changes, and/or extreme imaging artifact. Three examinations identified with use of natural language processing were excluded for importation issues. This results in a final development set of 274 non-consecutive contrast-enhanced CT examinations acquired as 5 mm thick axial slices in 251 patients from 1/1/2012 to 3/7/2017. This development dataset was intentionally enriched with CT examinations that contained adrenal masses to increase the proportion of adrenal masses in the training dataset above the relatively low population incidence of $4-6\%$. This development cohort included 433 normal adrenal glands (208 left and 225 right) and 104 masses (59 left and 45 right). Eleven glands were surgically absent (7 left and 4 right), and 12 patients had bilateral adrenal masses. The 274 CT examinations were divided into a training set of 214 CT examinations (78.1%), validation set of 25 CT examinations (9.1%), test set of 25 CT examinations (9.1%), and separate interreader test set of 10 CT examinations (3.6%) (figure 3-2). This patient cohort was drawn from a prior machine learning study performed to evaluate a different set of parameters [158]. To ensure that there was no artificial inflation of performance metrics due to data leakage, the set was split at the patient level. Specifically, if a patient had more than one examination, they were included only in a single data set (i.e., the training, validation, or test data sets) and not multiple data sets.

A total of 1066 consecutively acquired examinations performed at MGB from 11/1/2019 to 12/31/2019 were extracted for use as the secondary test set. Fifty examinations were subsequently excluded for incorrect axial section thickness (other than 5 mm) or dual-energy CT technique. This set consisted of 427 males and 589 females, with a median age of 62 years (IQR $50-72$ years). In this set, there were 1,951 normal adrenal glands (967 left and 984 right), and 76 masses (45 left and 31 right). Five glands were surgically absent (4 left and 1 right), and 9 patients had bilateral adrenal masses. In the resulting data set, 25 examinations were performed in duplicate patients, so the second examination was deleted, resulting in 991 examinations in

991 unique patients. As this set was not enriched with adrenal masses, incidence of adrenal masses was 3.8%, compared with 19.1% in the development set.

Table 3.1: Patient demographics and adrenal mass parameters in the development set and secondary test set.

Characteristic	Development Training Set	Development Validation Set	Development Test Set	Development Interreader Set	Secondary Test Set
Patient Characteristics					
No. of patients	196	23	24	8	991
No. of women*	105 (54)	14 (61)	10 (42)	4 (50)	578 (58)
Median age†	61 (52 - 69)	59 (51 - 73)	64 (53 - 75)	54 (46 - 64)	62 (72)
Examination Characteristics					
No. of CT examinations	214	25	25	10	991
No. of normal adrenal glands	343	37	37	16	1902
No. of adrenal masses	78	11	12	3	75
No. of adrenal resections	7	2	1	1	5
Gland Characteristics					
Left mass	46	6	6	1	44
Left normal	164	18	18	8	943
Left resected	4	1	1	1	4
Right mass	32	5	6	2	31
Right normal	179	19	19	8	959
Right resected	3	1	0	0	1

* Data in parantheses are percentages.

† Data in parantheses are IQRs.

A final tertiary dataset of 3064 consecutively acquired examinations (with no exclusions) was extracted. This dataset was also not enriched with adrenal masses, resulting in an incidence rate of 4.5%.

3.3.2 Segmentation Results

On the development test set of 25 CT examinations, the median DSC was 0.81 (IQR, 0.76 – 0.90) for left masses and 0.80 (IQR, 0.78 – 0.88) for left normal glands. The median DSC was 0.85 (IQR, 0.83 – 0.88) for right masses and 0.82 (IQR, 0.78 – 0.89) for right normal glands. Combining left and right, the median model DSC was 0.80 (IQR, 0.78 – 0.89) for the 37 normal glands and 0.84 (IQR, 0.79 – 0.90) for the 12 adrenal masses. Among the 10 CT examinations segmented by all five radiologists, the segmentations created by the different radiologists were shown to be similar. Annotations of an example case are shown in figure 3-7. Median interreader DSC ranged from 0.77 to 0.94 for all CT examinations (table 3.2), with a median of 0.89 (IQR, $0.78 - 0.93$) for normal adrenal glands and 0.89 (IQR, $0.85 - 0.97$) for adrenal masses. Median model-reader DSC ranged from 0.81 to 0.88 for all scans, with a median of 0.87 (IQR, $0.82-0.89$) for normal adrenal glands and 0.85 (IQR, $0.85-0.93$) for adrenal masses. The interreader DSC was not different than the model-reader DSC ($P = .35$), indicating that the segmentation performance of the machine learning algorithm did not differ significantly from that of the radiologists.

We note rater specific bias for segmentation on this development interreader dataset. Specifically, we observe that certain raters tend to either always under- or over-segment relative to the others. In figure 3-8, we see that rater W.W.M-S. generally under-segments while rater C.R.W. generally over-segments. We hypothesize that differences in experience level and training between raters can lead to such internal biases. We also observe that the model segments near the mean of the group, indicating that it learned to perform within the bounds of radiologist ground truth.

Table 3.2: Median interreader and reader-model DSCs for 19 adrenal glands in the development interreader set.

Reader No.	Reader 2	Reader 3	Reader 4	Reader 5	Model
Reader 1	0.89 ($0.58 - 0.98$)	0.82 ($0.67 - 0.97$)	0.91 ($0.80 - 0.97$)	0.90 ($0.75 - 0.99$)	0.87 ($0.77 - 0.94$)
Reader 2	1	0.77 ($0.52 - 1.00$)	0.94 ($0.59 - 0.99$)	0.91 ($0.61 - 1.00$)	0.87 ($0.55 - 0.93$)
Reader 3		1	0.80 ($0.67 - 0.99$)	0.79 ($0.60 - 0.99$)	0.81 ($0.74 - 0.93$)
Reader 4			1	0.92 ($0.74 - 0.99$)	0.88 ($0.74 - 0.93$)
Reader 5				1	0.88 ($0.66 - 0.93$)

Note — Data are DSCs, with minimum and maximum values in parentheses.

3.3.3 Classification Results

With use of the enriched development test set of 25 CT examinations, the optimal binarization threshold (the threshold at which the distinction between mass and no

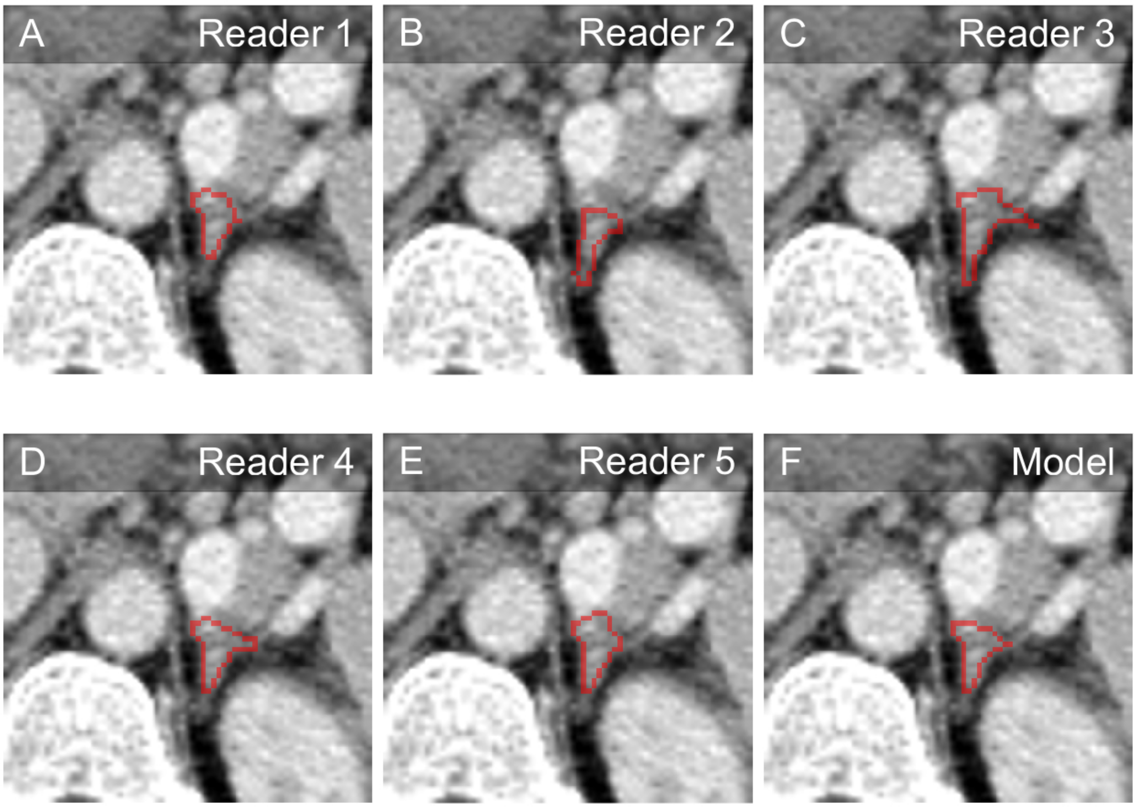


Figure 3-7: **Example interreader variability in adrenal gland segmentation compared with the model.** Contrast-enhanced axial CT images show the normal left adrenal gland of a 45-year-old woman. (A-E) Segmentations of the same adrenal gland as created by the five radiologists. (F) Segmentation created by the model. Some variability was observed across the segmentations, with reader 3 producing the largest segmentation and reader 2 the smallest.

mass is made) was determined to be 0.189, which was the operating point at which the model made the fewest errors in predicting if a gland was normal or mass-containing. At this threshold, the model correctly classified four of six left masses, 15 of 18 left normal glands, six of six right masses, and 18 of 19 right normal glands. In total, 10 of 12 masses were correctly classified, and 33 of 37 normal glands were correctly classified, yielding an overall sensitivity and specificity of 83% (95% CI: 55, 95) and 89% (95% CI: 75, 96), respectively. The area under receiver operating characteristic curve (AUROC) was 0.94.

AUROC and area under precision-recall curve (AUPRC) metrics for the secondary

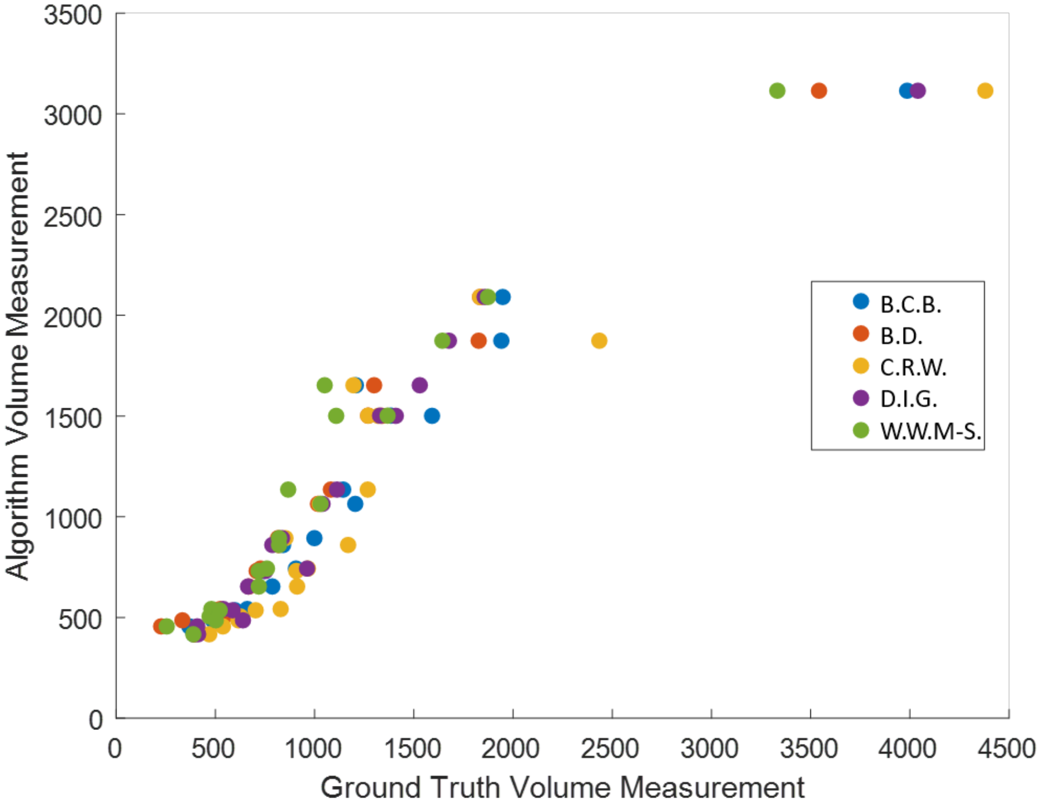


Figure 3-8: **Radiologists exhibit internally consistent biases for segmentation of adrenal glands.** This scatter plot shows five expert radiologists segmenting the same set of 19 adrenal glands. Raters exhibit internal consistency in under- and over-segmentation, with B.D. and W.W.M-S. having the smallest segmentation volumes and B.C.B and C.R.W having the largest. D.I.G. and the model both produced segmentations with volumes near the mean.

test set were 0.89 and 0.41, respectively (figure 3-9). Using the aforementioned optimal binarization threshold for the enriched development set on the secondary test set, the model correctly classified 33 of 44 left masses, 810 of 943 left normal glands, 19 of 31 right masses, and 924 of 959 right normal glands. In total, 52 of 75 masses were correctly classified, and 1734 of 1902 normal glands were correctly classified, yielding an overall sensitivity and specificity of 69% (95% CI: 58, 79) and 91% (95% CI: 90, 92), respectively. The κ was 0.31, indicating fair agreement. Model performance metrics at different operating points are presented in table 3.3, with the observation that as the binarization threshold approached 0.0, the model sensitivity became higher at the expense of specificity and model-reader agreement. Examples of model segmentation and classification successes and failures are shown in figures 3-10 and 3-11.

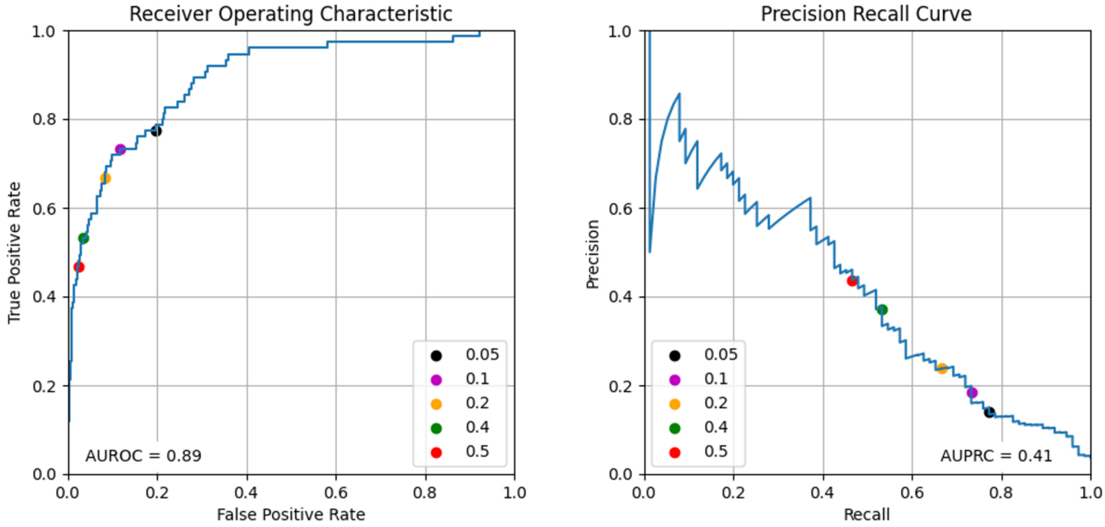


Figure 3-9: **Receiver operating characteristic curve and precision-recall curve for the secondary test set of 991 CT examinations.** The dots indicate the thresholds chosen to showcase sensitivity-specificity trade-off for our classification model. AUPRC = area under the precision-recall curve, AUROC = area under the receiver operating characteristic curve.

Table 3.3: Classification model performance on the secondary test set of 991 CT examinations with use of varying thresholds.

Threshold	Sensitivity (%)	Specificity (%)	Precision	Cohen κ
0.050	77 (58/75)	81 (1546/1902)	0.14	0.18
0.100	73 (55/75)	87 (1661/1902)	0.19	0.25
0.150	72 (54/75)	90 (1703/1902)	0.21	0.29
0.189	69 (52/75)	91 (1734/1902)	0.24	0.31
0.250	60 (45/75)	93 (1777/1902)	0.26	0.33
0.300	59 (44/75)	95 (1800/1902)	0.30	0.37

Note — Data in parentheses are number of CT examinations.

The results comparing the radiologist classification of masses with the automatic diameter measurements of the same 274 examinations using radiologist ground truth segmentations is shown in figure 3-12. These results show that automatic diameter measurements from the radiologist segmentations measured a mass ($\leq$ 10mm) in 119

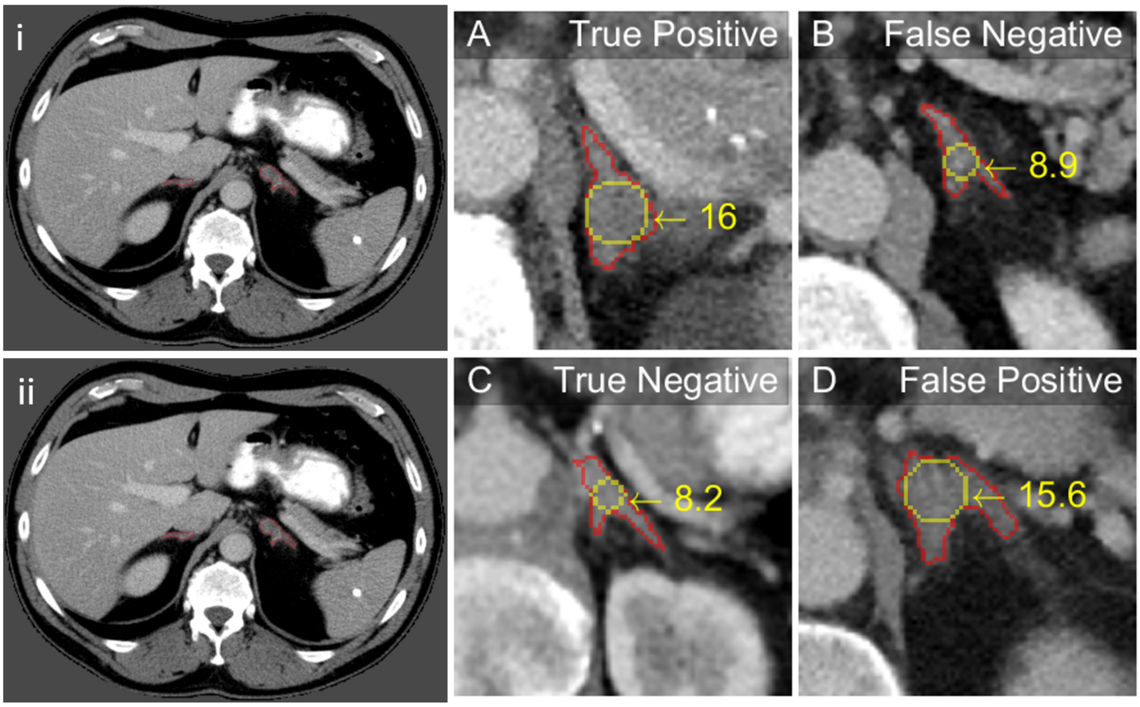


Figure 3-10: **Segmentation and classification examples.** (A-B) Segmentation example shows contrast-enhanced axial CT images in a patient from the development test set with a normal right adrenal gland and a left adrenal mass. (A) shows the ground truth radiologist segmentation, and (B) shows the model segmentation prediction of the glands. For this examination, the model achieved a Dice score of 0.90 for the normal right gland and 0.94 for the left adrenal mass. (C-F) Classification example shows contrast-enhanced axial CT images of four different left adrenal glands classified by a radiologist: two normal and two containing a mass. The red outline demonstrates the model segmentation, and the yellow circle represents an automated diameter measurement in millimeters of the potential mass. The segmentations in (C) and (E) show correct model inferences for a mass and a normal gland, respectively. The segmentation in (D) shows a false-negative result wherein the model predicted no mass, but the radiologist classified it as a mass. The segmentation in F shows a false-positive result wherein the model predicted a mass, but the radiologist classified it as no mass.

of 414 adrenal glands (29%) classified as normal by the radiologists. Conversely, two adrenal glands classified as having a mass by the radiologists did not exceed 10 mm according to the radiologist segmentations.

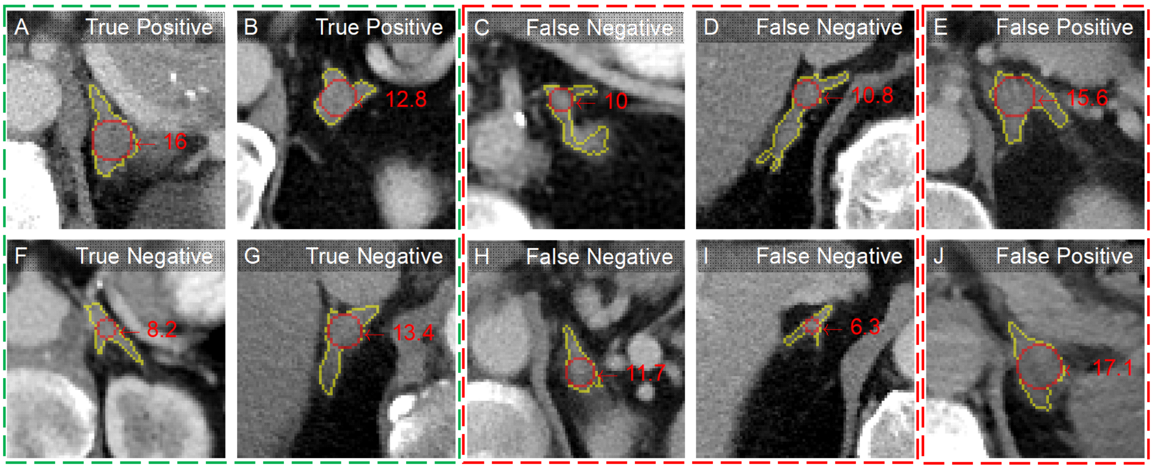


Figure 3-11: **Extra examples of model segmentation and classification.** (A-B) show examples of true positives. (F-G) show examples of true negatives. (C-D, H-I) show examples of false negatives. (E, J) show examples of false positives.

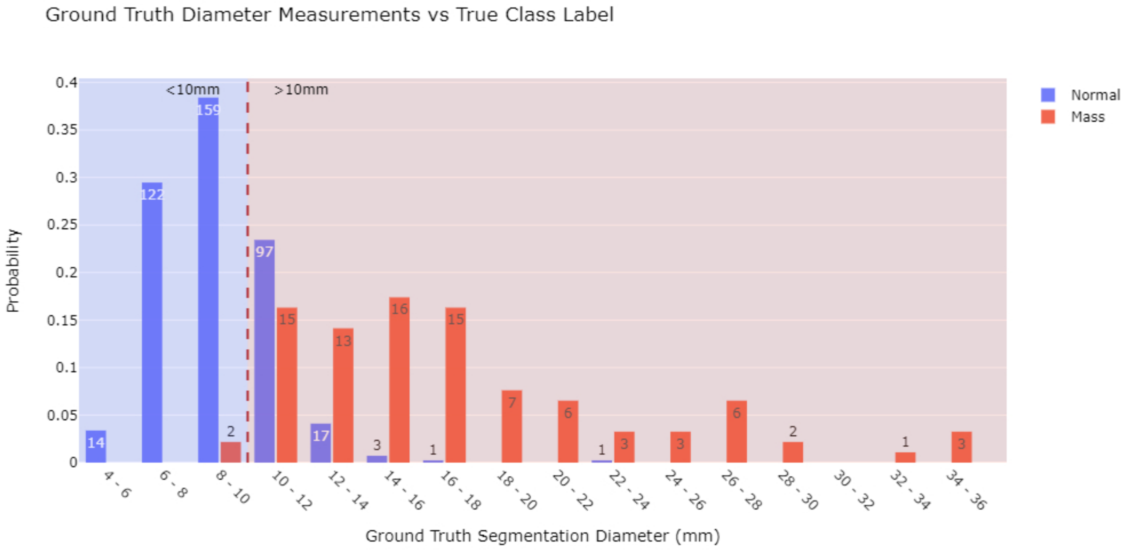


Figure 3-12: **True diameter measurements of normals and masses.** Bar graph shows ground truth diameter measurement versus true class label (normal vs adrenal mass) defined by the radiologists and distribution of adrenal glands according to the automated diameter measurements. Adrenal glands that were labeled as normal by the radiologists are in blue, whereas those labeled as containing a mass by the radiologists are in red. Note the overlap of the distributions, suggesting that classification of adrenal mass by size alone is inadequate.

3.3.4 Interreader Analysis

We further assess interreader variability for segmentation and classification on a small subset of the primary dataset. In particular, raters W.W.M-S. (who is an attending radiologist) and C.R.W. (who was a resident when performing this study for us) both segmented and classified 42 examinations (84 adrenals). For this subset, we note significant variability in both segmentation volumes ($p<0.001$) and detection of adrenal masses ($p<0.001$), with rater C.R.W. over-segmenting and over-calling masses compared to rater W.W.M-S (figure 3-13). Two example adrenal glands are shown in figure 3-14. Here, rater C.R.W. (in blue) over-segments compared to rater W.W.M-S. (in orange). Moreoever, C.R.W. classifies both as adrenal masses compared to W.W.M-S. who classifies them as normal. This is a well-known phenomena in medicine, where newer doctors tend to make more mistakes than more established and experienced doctors [170].

The sensitivity for mass detection on the secondary dataset is relatively low (only 69%). Our ground truth for this dataset was generated by two attending radiologists who interpreted the imaging specifically for this study. As all our data comes from clinical imaging at MGB, these images were already interpreted by a radiologist in the past, with the prior radiologist's read of the image stored in a structured patient radiology report. We manually collected this prior read of the image in order to compare it with our fresh read of the image. We find that there is significant variability in the reporting of adrenal masses ($p<0.001$). In total, either one or both radiologists classified a gland as a mass in 104 of 1982 glands. However, both radiologists only agreed in 53 of 104 (51%) of cases that a mass was present. In the remaining 51 of 104 (49%) of cases, only one of the two radiologists reported the presence of an adrenal mass. Examples of disputed adrenals are shown in figure 3-15.

3.3.5 Tertiary Dataset Analysis

There is significant cost involved in manually curating ground truth class labels for large-scale datasets, especially in the setting of low-prevalence diseases. In routine

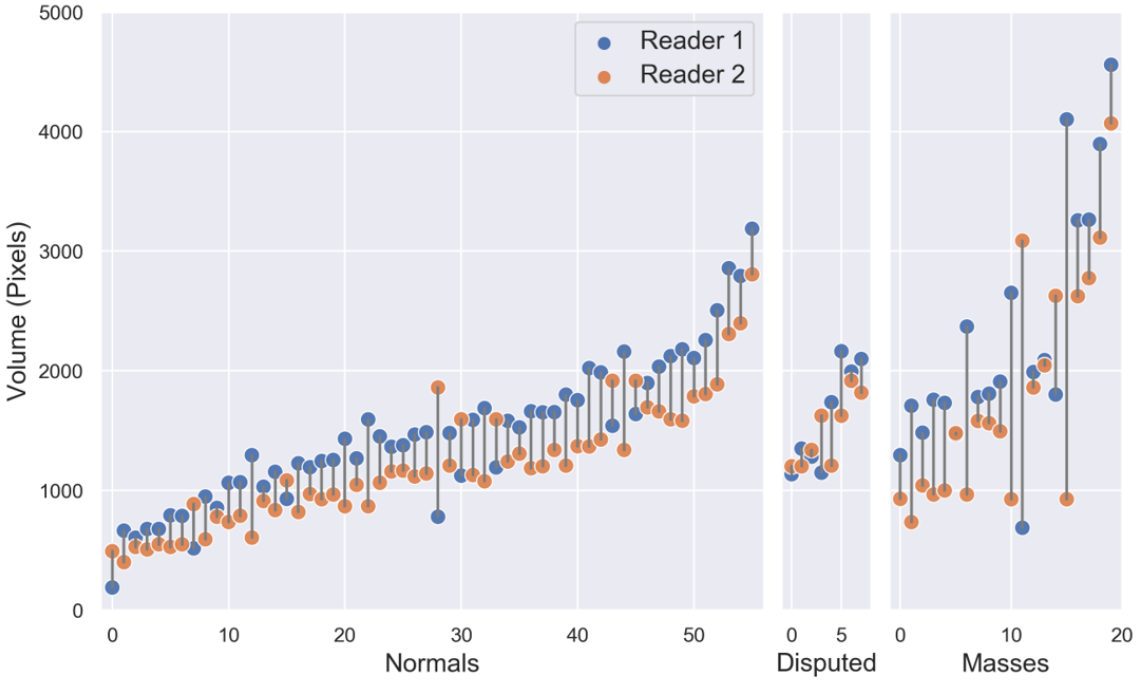


Figure 3-13: **Interreader variability in segmentation and classification of adrenal glands.** Two readers segmented and classified a set of 84 adrenals. Reader 1 (in blue) is a resident and reader 2 (in orange) is an attending doctor. This plot shows the volume of the segmented adrenals and the classification provided by the two readers. Reader 1 over-called 8 masses compared to reader 2, indicating substantial variability for the detection of adrenal masses.

clinical care, radiologists interpret medical imaging and write findings into a structured radiology report. In lieu of manually extracting class labels from radiology reports (or making a fresh read of the imaging), natural language processing (NLP) algorithms trained to read reports can be used to cheaply provide weak class labels. For our tertiary dataset, non-resected adrenal glands were weakly classified as normal or mass by the AdrenoBERT NLP model. If there was a discrepancy in the class label between the NLP read of the radiology report and the prediction by our classification model, an expert radiologist adjudicated the case, verifying that the AdrenoBERT correctly interpreted the report as well as making a fresh read of the CT. This fresh, adjudicated read was used as the gold-standard ground truth.

The AdrenoBERT and image classification models produced discrepant results on 686 of 6128 (11.2%) adrenals, which were henceforth adjudicate manually. Against

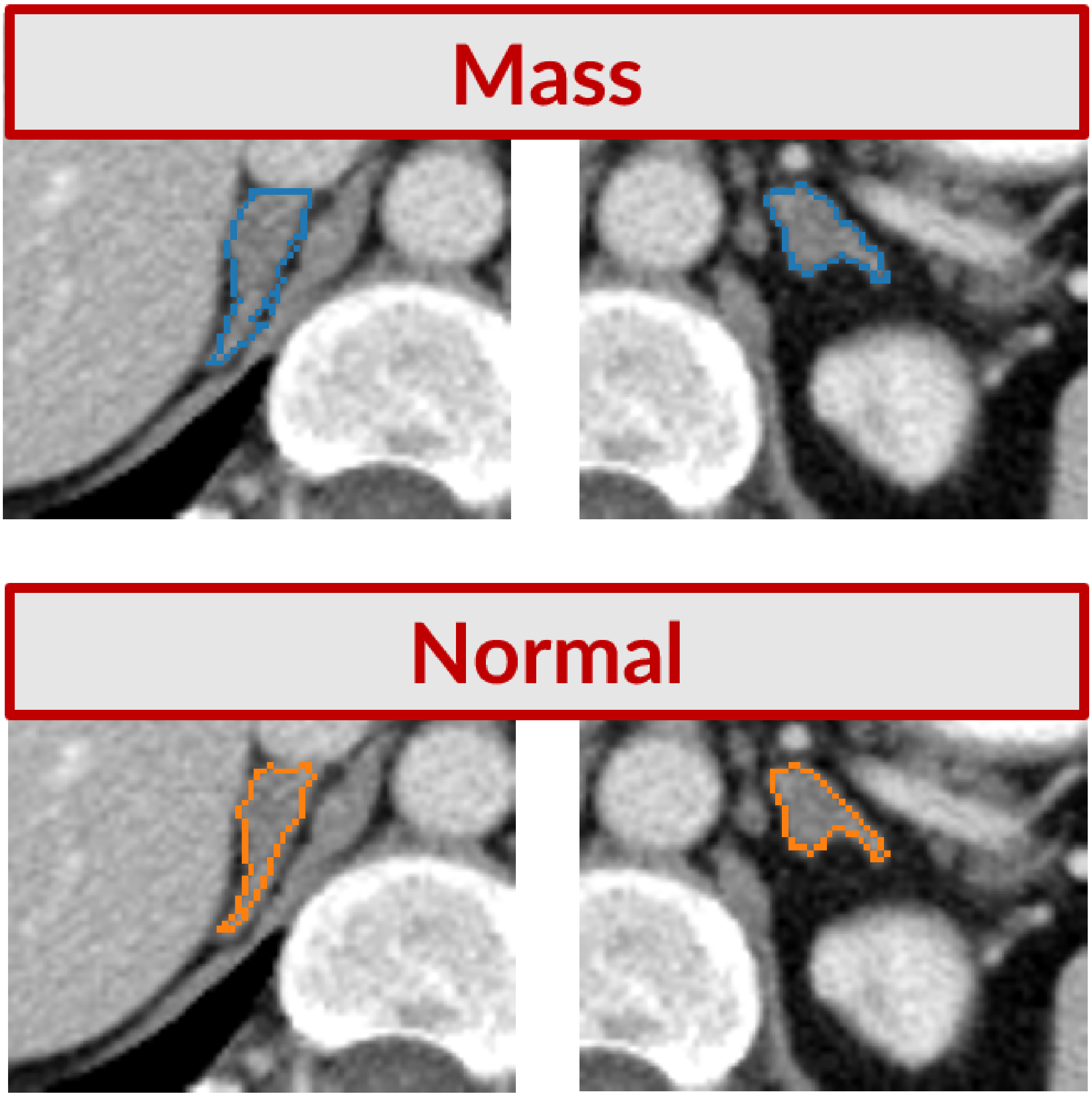


Figure 3-14: **Example of interreader variability between two expert radiologists.** Reader 1 (in blue) is a resident and reader 2 (in orange) is an attending doctor. Reader 1 over-segments these two adrenals compared to reader 2. Moreover, reader 1 calls both of these adrenals as mass whereas reader 2 calls both of these as normal.

the gold-standard, our imaging classification model had a sensitivity, specificity, and precision of 86%, 90%, and 30%, respectively. We also report an AUROC of 0.93 and an AUPRC of 0.65, visualized in figure 3-16.

Against the original radiology report, we report rates of 81%, 90%, and 25%, respectively. The difference in reported sensitivity and precision is indicative of significant interreader variability for the task of adrenal gland classification ($p<0.001$).

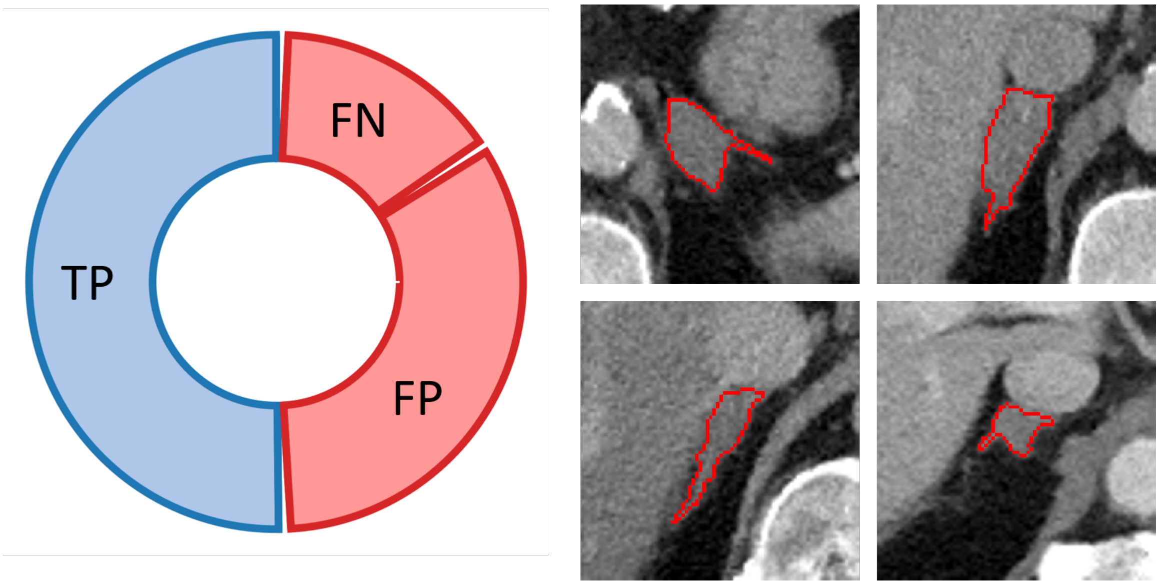


Figure 3-15: **Interreader variability for the classification of adrenal masses.** Detecting adrenal masses in clinical practice can be challenging. We note only 51% agreement for the presence of an adrenal mass. Examples of disputed masses is shown in the right panel of the figure.

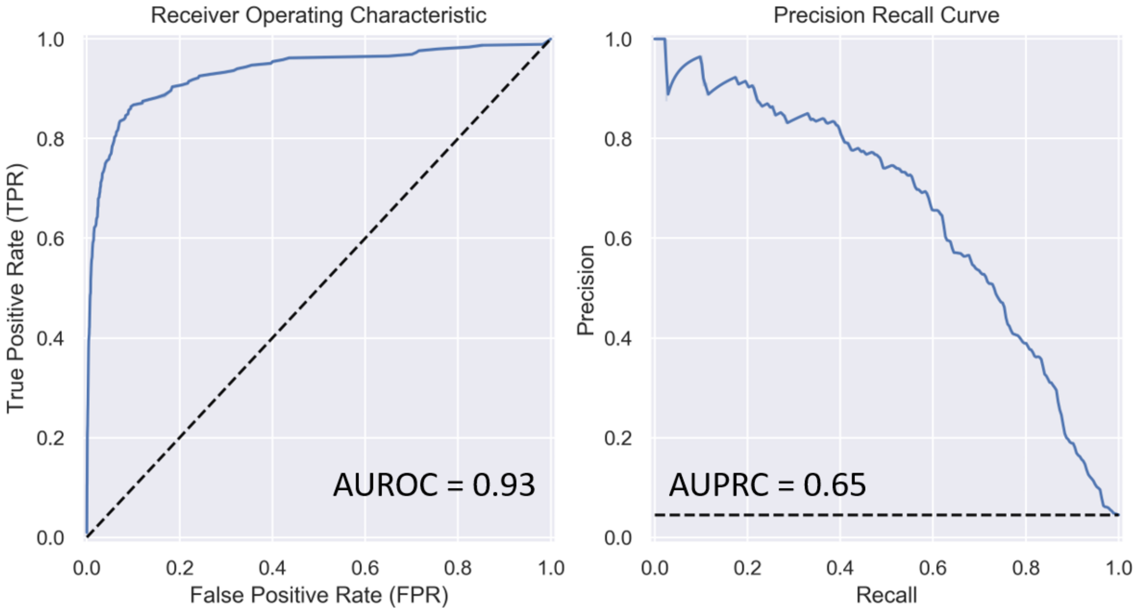


Figure 3-16: **ROC and PRC for the tertiary dataset.** Against the gold standard, our image classification model shows an AUROC and AUPRC of 0.93 and 0.65, respectively.

Against weak NLP labels, we report rates of 71%, 90%, and 23%, respectively. The estimate of imaging model performance when using weak NLP labels is moderately lower than when using gold-standard labels, indicating that there is scope to refine the AdrenoBERT model in future studies.

3.4 Discussion

Adrenal masses are common, but radiology reporting and recommendations for management can be variable. This study demonstrated that a machine learning algorithm can accurately segment adrenal glands and differentiate between normal glands and those containing masses. The segmentation model was able to reliably segment images of the adrenal glands, with performance similar to manual segmentations from radiologists, as evaluated with use of the Dice similarity coefficient (DSC). There was moderate interreader agreement between the five radiologists performing segmentations, with DSC ranging from 0.77 to 0.94 on the same 19 adrenal glands. The model DSC fit within this variation, ranging from 0.81 to 0.88.

For the classification task of differentiating normal adrenal glands from masses, the model reached a sensitivity of 69% and a specificity of 91% on the secondary test set of 991 CT examinations. Qualitative assessment of the glands that were not classified correctly revealed that most of the errors were due to a failure of the classification model (presence or absence of an adrenal mass) rather than failure in the segmentation model (outlining the adrenal gland). For example, only six of 23 false-negative results (algorithm reports no mass, but the radiologist classified an adrenal mass) in the consecutive test were attributed to segmentation failures (i.e., the adrenal mass was not correctly outlined, resulting in downstream classification errors).

Comparison of radiologist classification of a mass with the automatic diameter measurements shows the nuance and variability of radiologists defining an adrenal mass. These measurements were made on the same images, using radiologist manual segmentations as the reference standard. A total of 119 of 414 examinations (29%) classified as normal by the radiologists had diameters of 10mm or greater, highlighting the complexity in determining what is and is not a mass, and that solely relying on a deterministic 10mm diameter decision boundary is insufficient. Alternatively, it is possible that radiologists themselves demonstrate great variability in defining masses and could benefit by having assistance from a more standardized algorithm [171]. Given how much variability exists between radiologists in differentiating small masses from normal glands, it is not surprising that our classification model produced errors even in instances of highly accurate automatic segmentation.

Our results are promising given that the adrenal gland is an inherently challenging organ to segment compared with larger organs (such as the liver and kidneys), as these glands are small and change in position due to respiration, and their shape, size, and location can vary by laterality and patient. In addition, adrenal glands have soft-tissue attenuation at CT that is similar to that of adjacent structures, including the liver, pancreas, kidneys, and vasculature. In patients with a paucity of intra-abdominal fat, the adrenal glands may be even more challenging to delineate from adjacent structures without the contrasting fat around the glands to separate them from other structures. Finally, there may be external mass effect on the glands, for example from an adjacent liver cyst or mass.

There have been several prior studies investigating the ability of a machine learning algorithm to automatically segment adrenal masses [133, 155, 172, 173]. Saiprasad et al. [155] reported an initial effort to use random forest classification to detect the adrenal glands and determine if an adrenal abnormality was present. This small study consisted of 20 adrenal glands in 10 patients and demonstrated a 79.9% sensitivity and 99% specificity for classifying the adrenal glands compared with manual segmentation. Building upon this approach, Koyuncu et al. [133] used an automated pipeline in 32 adrenal tumors and achieved a sensitivity of 86% and a specificity of 99%. Both of these studies evaluated substantially smaller patient samples than our study and used different machine learning methods.

3.5 Limitations

Our study has limitations. First, this was a single-center retrospective study, and model performance at other sites or when the model is used in a prospective manner is uncertain. Second, images of adrenal glands in the development set were segmented by one of five different radiologists, including trainees, which may have led to increased variation in manual segmentation compared with expert annotation and/or consensus of multiple radiologist segmentations of the same scan. Third, we used an overall definition of an adrenal mass as being 10 mm or greater in the short axis. Although this is the accepted definition for incidental adrenal nodules, functional masses or metastases may be smaller than 10 mm. In figure 3-17, an example of two very similar adrenal glands is shown. However, the adrenal on the right is an example of a hyperfunctioning adrenal mass that is smaller than 10 mm. Our model was not trained to identify and classify such cases, thus accidently mis-classifying it as normal. Fourth, although we evaluated the performance of the model at the level of individual glands, adrenal glands are almost always paired structures. This may have introduced bias. Our study did not have histopathologic examination as a reference standard, and it is possible that what was considered a mass at imaging was actually normal and vice versa. Last, the impact of imaging characteristics, such as CT attenuation, on mass identification were not assessed.

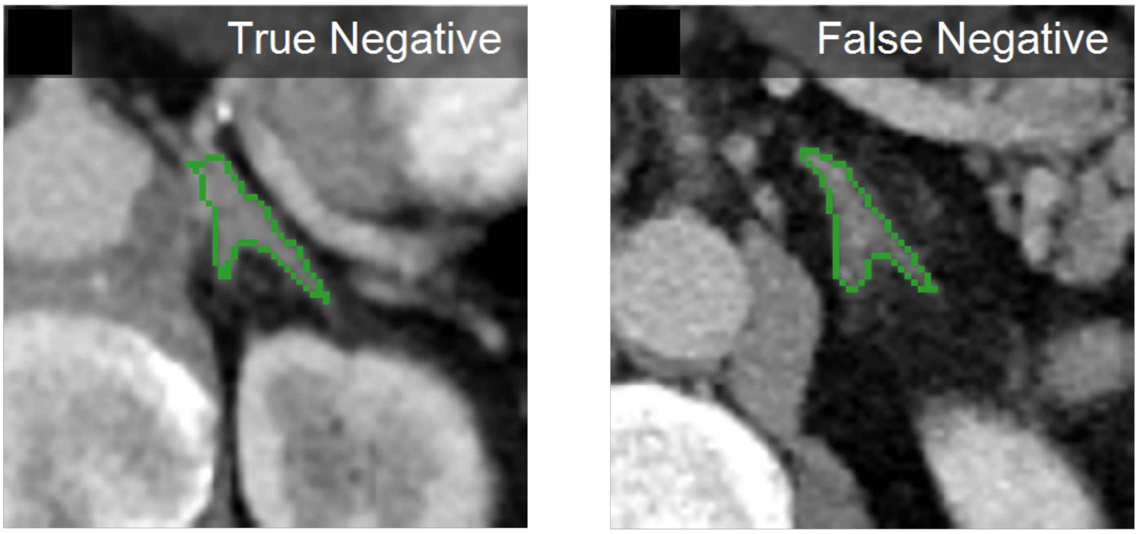


Figure 3-17: **Example of a small hyperfunctioning adrenal mass.** Our model was not trained to detect small hyperfunctioning adrenal masses, which can often be smaller than 10 mm. Even though both adrenals in the figure are similar, the adrenal on the left is normal while the adrenal on the right is hyperfunctioning.

3.6 Conclusion

In conclusion, we propose a two-stage machine learning pipeline to automatically segment the adrenal glands at contrast-enhanced CT and then classify the glands as normal or mass-containing. This tool may be used to assist radiologists in accurate and expedient image interpretation and potentially decrease interreader variability. Future work is needed to improve the classification stage of our model, as well as expand on the scope of the classification task by reviewing prior imaging and assessing for mass stability or growth.

Chapter 4

Metastatic Brain Tumor Segmentation and Longitudinal Tracking

Effective analysis of longitudinal MR is one of the backbones for assessing tumor response in patients with brain metastases. Currently, manual delineation of all metastases is not only prohibitively time-consuming, but subject to significant inter-rater variability. As a result, most radiographic response criteria, including the Response Assessment in Neuro-Oncology Brain Metastases (RANO-BM) criteria, restrict response assessment to a select number of target lesions, failing to capture the full extent of the disease burden and potentially missing longitudinal changes in non-target tumors. In this chapter, we discuss a deep learning based approach for brain metastases segmentation on MRI, and validate model performance across lesion sizes. Moreover, we automate standard response assessment classification and quantify model uncertainty, highlighting a potential method to automatically identify false positives and other mistakes.

4.1 Introduction

Brain metastases (BM) are the most common form of intracranial tumors in adults, affecting around 20% of all cancer patients [174, 175, 176]. This number is expected to increase as systemic treatments for primary tumors improve [134, 177, 178]. Even following standard treatment regimen of surgery and/or stereotactic radiation, median survival post diagnosis of BM ranges from 2.7 to 24 months [134]. The current standard to determine treatment response and assess tumor progression in clinical trials is the Response Assessment in Neuro-Oncology Brain Metastases (RANO-BM) criteria, which is based on (i) uni-dimensional measurements of a specific number of target lesions (enhancing lesions with diameter $\leq$ 10 mm), and (ii) qualitative assessment of non-target lesions from contrast-enhanced magnetic resonance imaging (MRI) [179]. However, the manual determination of tumor boundaries needed for uni-dimensional measurements can be challenging when there is heterogeneous contrast enhancement, tumor boundaries are diffuse, or when contrast relative to surrounding normal brain parenchyma is blunted due to treatment effects, resulting in significant inter- and intra-rater variability [180]. As such, there has been interest in developing automated methods for segmentation and response assessment to improve reproducibility and reduce provider burden in performing RANO-BM measurements.

Although uni-dimensional measurements are currently used as a measure of tumor burden, they may represent an incomplete measure when tumors are irregularly shaped. Furthermore, because only qualitative evaluation is made of non-target lesions, there is inherent subjectivity in their assessment. Smaller, non-target lesions may be neglected or missed due to the lack of quantitative emphasis. Thus, volumetric assessment of both target and non-target lesions represents a more comprehensive measure of tumor burden. This is reflected in a recent consensus paper on brain tumor imaging in clinical trials, which noted volumetric analysis as an important mode of improvement [181]. However, this has not been adopted in routine practice due to the labor intensive task of manual segmentation, especially when brain metastases are numerous. Indeed, 47% of patients have more than one BM, with 41% of patients having 4 or more metastases [182, 183]. Examples of brain metastases of different sizes and number is shown in figure 4-1. An automated tool for longitudinal volumetric tracking of brain metastases that is integrated into the radiographic analysis pipeline could help facilitate the use of tumor volume as a response endpoint in clinical trials. Furthermore, integration into the clinical workflow can assist physicians with real-time treatment decision-making.

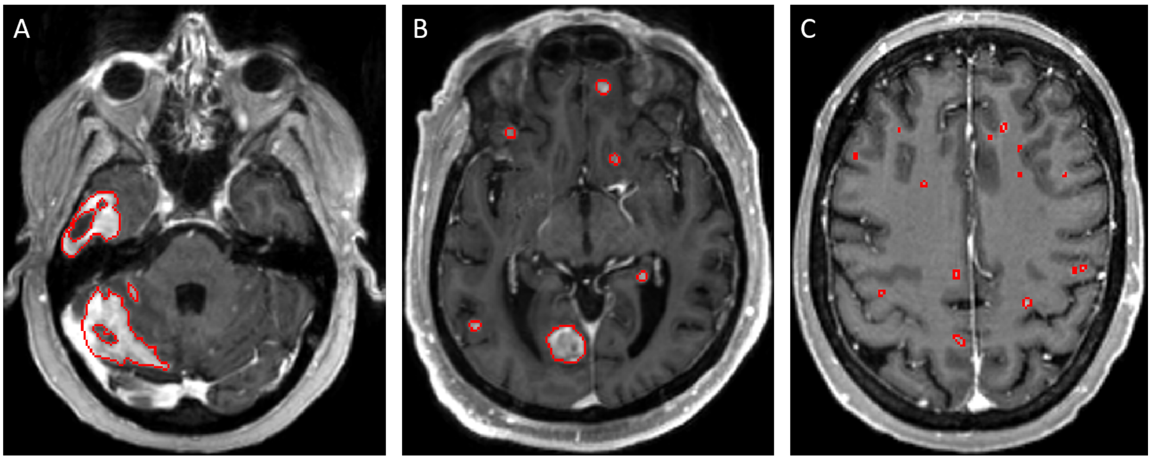


Figure 4-1: **Examples of brain metastases.** Brain metastases can present with a wide range of shapes and sizes. Panels A, B, and C show examples of large, medium, and small metastases, respectively.

With advances in machine learning techniques, deep learning has become the state-of-the-art approach for lesion segmentation within medical imaging [184, 185, 186, 100]. Recent work has shown the potential of deep learning for volumetric response assessment in primary gliomas [187, 188] and other studies have used deep learning to automatically segment brain metastases [140, 189, 190, 191, 192, 193, 194]. A major limitation of these prior studies on brain metastases is that they were performed on a single patient visit without comparison to subsequent visits. In the setting of single tumors, this approach works but in the setting of multiple BM, being able to track multiple unique tumors longitudinally is critical.

In this study, we developed a pipeline for automatic segmentation of brain tumor metastases, with validation on two independent patient cohorts. Notably, we modify the neural network loss function to emphasize tumor boundaries, improving segmentation performance. Using an ensemble of networks, we also derive voxel-wise and lesion-wise uncertainties, enabling us to identify potential model mistakes. We also automate the quantification of measurable tumor burden as given by the RANO-BM criteria. Finally, we developed an algorithm to perform longitudinal tracking of individual lesions and automatic growth rate characterization.

4.2 Materials and Methods

4.2.1 Primary and Secondary Patient Cohorts

This study was conducted following approval from the Partners Institutional Review Board. All patients met the following criteria: i) histopathologically or clinically confirmed BM and ii) available MPRAGE-post imaging sequence with mild to no motion/ringing artifacts. A primary dataset was constructed from a set of 46 patients (118 timepoints total) participating in a clinical trial of pembrolizumab from Massachusetts General Hospital (MGH) (NIH clinical trial ID: NCT02886585 [195]) and 36 patients (64 timepoints) participating in a clinical trial of cabozantinib from MGH (NIH clinical trial ID: NCT02260531 [196]). Pembrolizumab patients were used as the training set, with cabozantinib patients equally split into validation and testing cohorts. A secondary dataset comprised of 148 patients (885 timepoints total) from Brigham and Women's Hospital (BWH) who were undergoing stereotactic radiosurgery treatment was retrospectively acquired from April 2004 to November 2014. This dataset was not used for model development and served as an independent testing set. All patient examinations were viewed and annotated in Slicer3D [197]. For the primary dataset, manual segmentations of the contrast-enhancing lesions were performed by a board-certified neuro-oncologist with 10+ years' experience. For the secondary dataset, segmentations were first manually segmented by a neuro-oncologist and then manually edited by a board-certified neuro-radiologist with 16 years' experience. To better understand how model performance varied as a function of metastasis volume, each dataset was sub-divided into groups of consisting of small ($< 125\text{mm}^3$), medium ($\geq 125\text{mm}^3$ and $< 1000\text{mm}^3$), and large ($\geq 1000\text{mm}^3$) metastases. To better understand how model performance varied as a function of number of metastases, each dataset was sub-divided into groups of consisting of few (< 4), multiple (≥ 4 and < 10), and many (≥ 10) metastases. Full dataset characteristics are found in Table 4.1.

Table 4.1: Patient demographics and metastatic lesion characteristics for primary and second cohorts.

Characteristic	Primary Training Set	Primary Validation Set	Primary Test Set	Secondary Test Set
Patient Characteristics				
No. of patients	46	18	18	148
No. of women*	40 (N/A)	18 (100)	18 (100)	94 (64)
Median age†	59 (N/A)	50 (N/A)	50 (N/A)	61 (54 - 68)
Examination Characteristics				
No. of MR examinations	118	32	32	885
No. of small lesions	604	154	258	3161
No. of medium lesions	385	172	247	1367
No. of large lesions	182	53	43	722
No. with few lesions	44	6	5	546
No. with multiple lesions	34	14	9	235
No. with many lesions	40	12	18	104
Primary Cancer Type				
Lung*	7 (N/A)	0 (0)	0 (0)	77 (52)
Breast*	20 (N/A)	18 (100)	18 (100)	28 (19)
Melanoma*	5 (N/A)	0 (0)	0 (0)	27 (18)
Gastrointestinal*	0 (N/A)	0 (0)	0 (0)	6 (4)
Renal*	1 (N/A)	0 (0)	0 (0)	5 (3)
Other/Unknown*	7 (N/A)	0 (0)	0 (0)	5 (3)

* Data in parantheses are percentages.

† Data in parantheses are IQRs.

(N/A) means data was not available.

4.2.2 Deep Learning Segmentation Algorithm

Our segmentation model is a 3D U-Net [99] which takes a high-resolution T1-weighted contrast-enhanced (T1-CE) sequence image as input and outputs a probability map of the likely brain metastases. Full details about the network architecture and training optimization is described below.

Pre-processing

To mitigate variability between patients stemming from potential differences in imaging protocols, we apply the following pre-processing steps. First, we resample all data to 1mm isotropic resolution and skull strip using ROBEX [198]. To compensate for intensity inhomogeneity, we apply N4 bias correction [199] and normalize the image intensities to have zero mean, unit variance (based on non-zero intensity voxels only). Finally, all volumes are tightly cropped to remove empty voxels (background intensity) outside the brain region.

Network Architecture

Guided by successful approaches from previous brain tumor segmentation work [101], we utilize a symmetrical 3D U-Net [99] architecture as the backbone for our model. A schematic of this architecture is shown in figure 4-2. We devise a 6 level U-Net which takes a T1-weighted contrast-enhanced (T1-CE) sequence image as input and outputs a probability map of the likely brain metastases. This probability map is then binarized to create a label map using a threshold of 0.5. To ensure our model learns both adequate anatomic context and rich representations under the constraints of GPU memory, we use 32 filters in the 3x3x3 convolutions in the first layer, and double this number of filters as we go deeper into the network. Feature map downsampling and upsampling is accomplished through strided convolution and trilinear interpolation, respectively. Instance Normalization [200] in lieu of Batch Normalization [164] is used in order to accommodate the smaller batch size necessary to train a large patch 3D model [101]. Rectified linear unit (ReLU) [163] activation was used in all layers, with the exception of the final sigmoid output. And to encourage faster convergence and ensure that deeper layers of the decoder are learning semantically useful features, we employ deep supervision [165] by integrating segmentation outputs from all but the two deepest levels of the network.

Loss Function

In order to maximize the Dice Similarity Coefficient (DSC) [71] between the ground truth label map p and predicted label map q, we use the following implementation of soft dice loss:

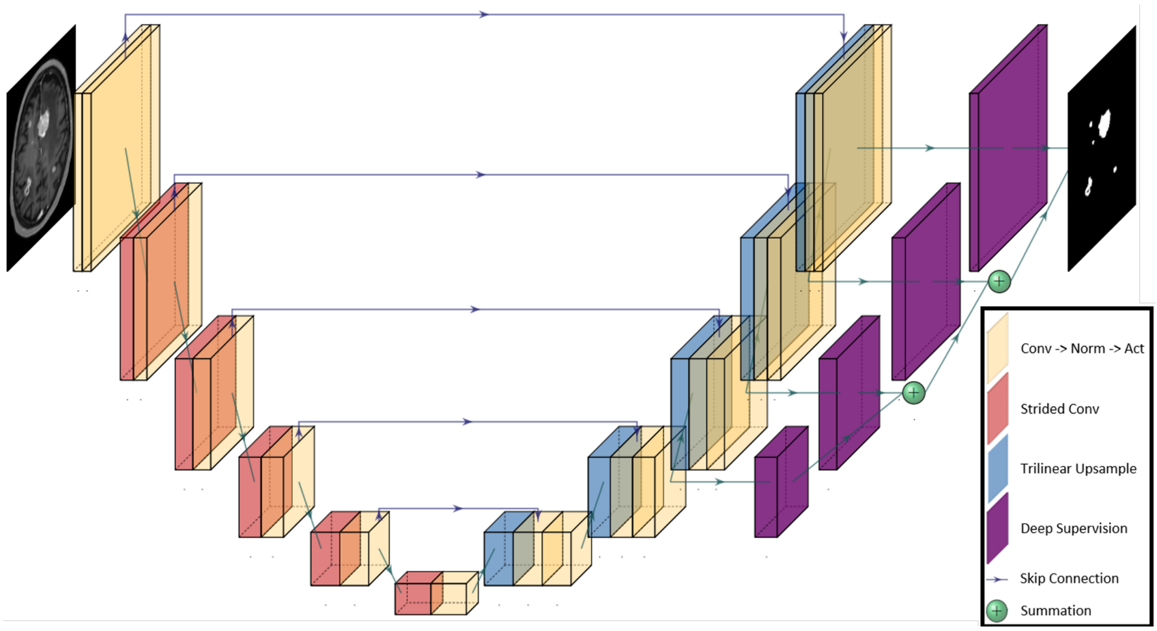


Figure 4-2: **Schematic of U-net architecture.** We devise a 6 level U-net which takes a T1-weighted contrast-enhanced (T1-CE) sequence image as input and outputs a probability map of the likely brain metastases.

$$\mathcal{L}_{DSC}(p,q) = 1 - \frac{2\sum pq + \epsilon}{\sum p + \sum q + \epsilon} \tag{4.1}$$

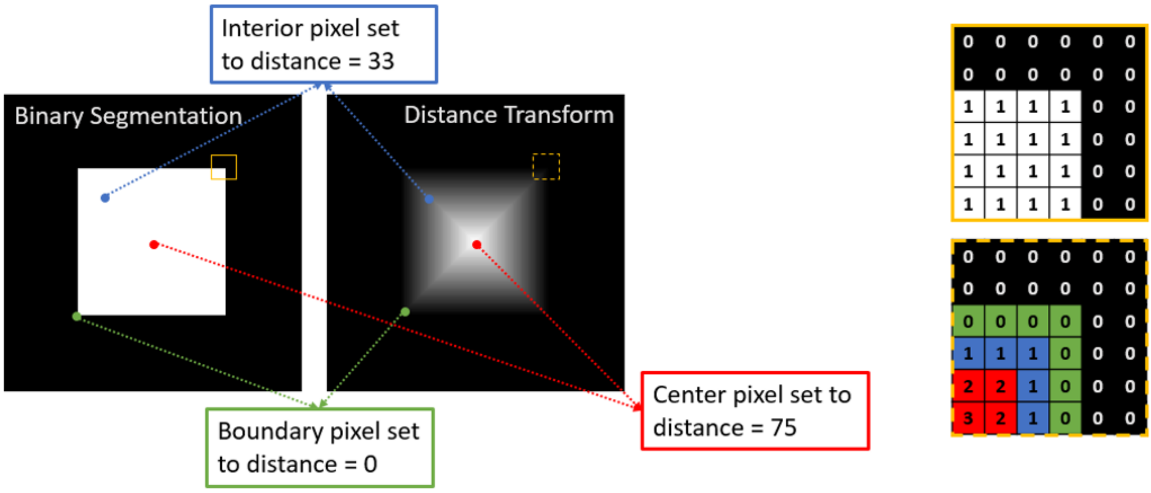
where ϵ is used to prevent floating point instability when the magnitude of the denominator is small (set to 1).

We complement the soft dice loss with cross-entropy loss, which we find enables more refined segmentation outputs. To handle the large class imbalance present in this segmentation task, we apply a boundary-reweighting term to the cross-entropy loss [201]. This reweighting map is created as follows. First, all ground truth label maps are binarized and converted into edge images. Next, the euclidean distance transform (a simple example of which is shown in figure 4-3) of these edge images is computed to produce a raw distance map, where the value at each voxel quantifies how far it is from the boundary of a metastasis. This map is subsequently inverted and rescaled such that voxels on the boundary are weighted 6 times more than voxels far away from the boundary. This ensures that easily classified voxels (such as those containing normal tissue distal to the tumor) are given less importance than the difficult to classify voxels (both foreground and background) near the peripheries of the metastases. An example of a reweighting map is shown in figure 4-4. This boundary-weighted cross-entropy loss $\mathcal{L}_{CE_{BW}}$ is implemented as follows:

$$\mathcal{L}_{CE_{BW}}(p, q) = -BW \sum p \log q \tag{4.2}$$

where BW is the boundary-reweighting map. The total loss for our network is the unweighted sum of the two losses:

$$\mathcal{L}_{total}(p, q) = \mathcal{L}_{DSC}(p, q) + \mathcal{L}_{CE_{BW}}(p, q) \tag{4.3}$$


Interior pixel set to distance = 33
Binary Segmentation
Distance Transform
Center pixel set to distance = 75
Boundary pixel set to distance = 0

0	0	0	0	0	0
0	0	0	0	0	0
1	1	1	1	0	0
1	1	1	1	0	0
1	1	1	1	0	0
1	1	1	1	0	0

0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	1	1	0	0	0
2	2	1	0	0	0
3	2	1	0	0	0

Figure 4-3: **Simple explanation of a one sided distance transform.** To convert a binary label map into a distance transform, each point inside the region of interest is given the value of the distance to the closest edge pixel. For instance, a point already on the edge is given a value of 0, whereas a point two pixels away from the edge is given a value of 2.

Optimization

We train our network on patches of size 128 × 128 × 128 voxels with batch size 1. As patches of this size cover most of the brain region already, we sample patches at random during training, seeing no performance change from using special sampling heuristics. Training is done using the SGD optimizer with decoupled weight decay

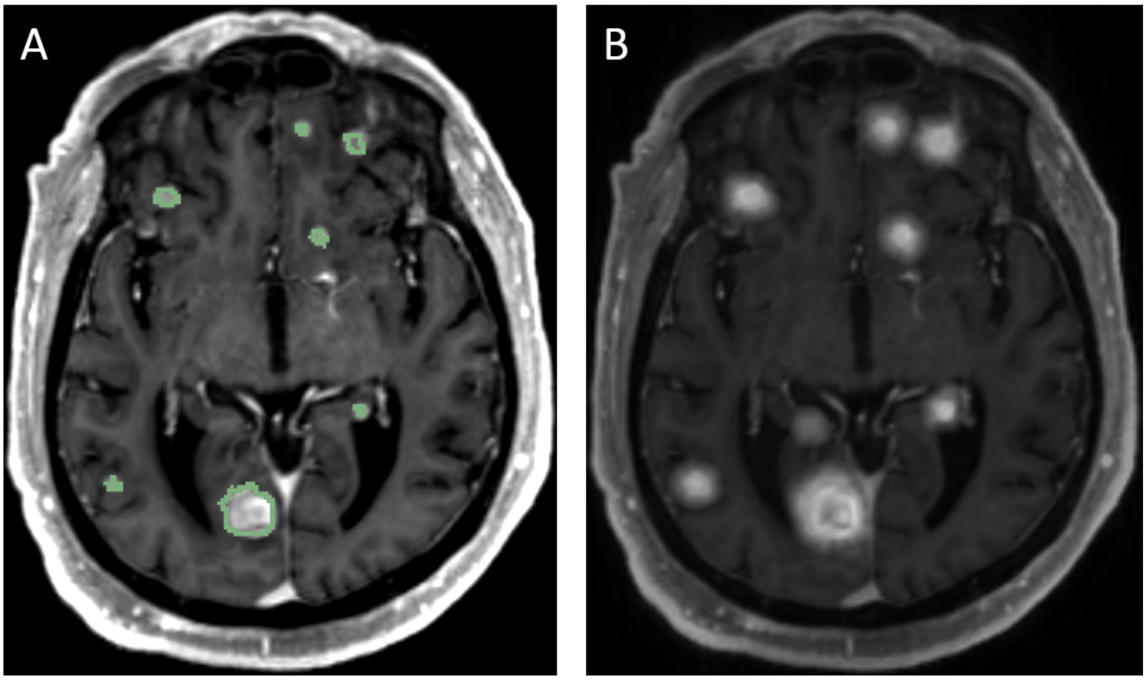


Figure 4-4: **Example of a reweighting map.** A) An MRI of a patient with ground truth segmentation overlaid. B) The reweighting map is overlaid onto the image, showing that the network is encouraged to focus on the boundaries of tumors more than other regions.

[56] and we progressively decrease the learning rate via the following cosine decay schedule:

$$\eta_t = \eta_{min} + 0.5(\eta_{max} - \eta_{min})(1 + cos(\pi T_{curr}/T)) \tag{4.4}$$

where η_{max} is our initial learning rate (set to 0.1), η_{min} is our final learning rate (set to 0.0001), T_{curr} is the current iteration counter, and T is the total number of iterations to train for (set to 150 epochs). To mitigate overfitting, we apply weight decay of 0.00002 to all convolutional kernel parameters, leaving biases and scales unregularized. Furthermore, we apply real-time data augmentation during the training process. Specifically, we utilize random mirror axis flips about all three axes along with anisotropic scaling (0.75 to 1.25), rotations ($-15°$ to $15°$), shearing (-0.15 to 0.15), and translations (-5 to 5 pixels). Intensity augmentation in the form of gamma correction (.75 to 1.25) is used as well. All augmentations are applied with probability

0.5.

Training a single model using the labeled primary training cohort of 118 examinations took around 10 hours on a NVIDIA Tesla V100 32GB GPU.

Inference

At test time, we pass the entire image into the network to be segmented in one pass, as opposed to using a sliding window of patch size 128 × 128 × 128. We find that this is both more efficient and also leads to better segmentation quality, presumably due to less edge effects stemming from the use of zero-padded convolutions. We apply simple test-time augmentation by averaging the results from 8 mirror axis-flipped versions of the input volumes. To further boost performance, we average the results from an ensemble of five models (all trained from scratch).

4.2.3 AutoRANO-BM

We adapt the AutoRANO algorithm [187] used to automatically derive RANO measurements for primary gliomas towards RANO-BM measurements for metastatic tumors. For each metastasis in the patient timepoint, the AutoRANO-BM algorithm searches for the axial slice with the largest lesion area and determines if the lesion is measurable. A measurable lesion is defined as having a minimum length of both perpendicular measurements $\geq$ 10mm. If the lesion is measurable, the maximum uni-directional diameter is automatically calculated. Finally, an automatic RANO-BM measure is derived by summing the maximum uni-directional diameters of up to the five largest measurable lesions. An example is shown in figure 4-5.

4.2.4 Response Assessment Classification

After AutoRANO-BM measures are computed for all timepoints, each timepoint following the baseline visit is categorized as progressive disease (PD), stable disease (SD), partial response (PR), or complete response (CR) as outlined by the RANO-BM working group [179]. Briefly, PD is categorized by a greater than 20% increase in

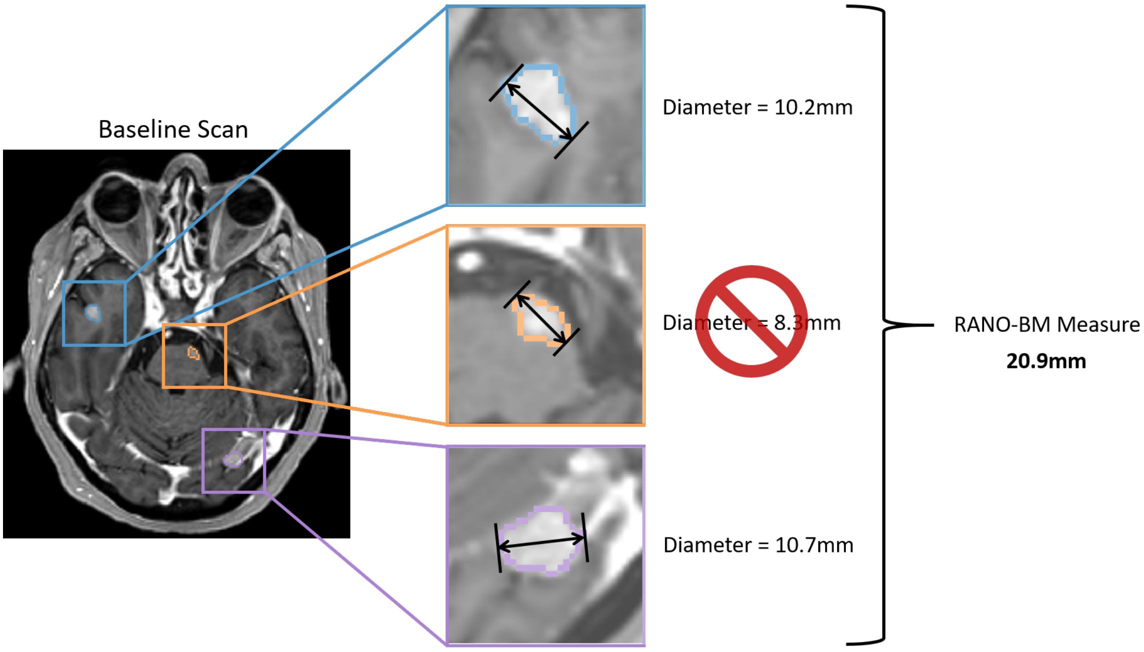


Figure 4-5: **AutoRANO-BM example for a brain metastases patient.** A patient presents with three distinct lesions (colored blue, orange, and purple). The maximal uni-dimensional diameter is automatically calculated for each metastasis. Since the orange lesion is too small (i.e. it has a diameter less than 10 mm), we ignore it per the RANO-BM criteria. The RANO-BM measure for this patient is then the sum of the diameters of the blue and purple metastases. We note that this example shows a simplification of the real procedure, which would be done in 3D for all metastases that this patient may have (not just those visible on this slice).

RANO-BM measure relative to the nadir timepoint. SD is categorized by a less than 20% increase relative to the nadir timepoint and a less than 30% decrease relative to the baseline. PR is categorized by a greater than 30% decrease relative to the baseline. And CR is categorized by complete disappearance of all lesions. An schematic presenting this is shown in figure 4-6.

4.2.5 Longitudinal Tracking

For a given patient, all timepoints are affinely co-registered to the baseline visit. Using connected components analysis, every metastasis in the baseline scan is given a unique numeric identifier. For a subsequent timepoint, every metastasis in that scan is cross-referenced with every prior scan to determine if the metastasis is unique or

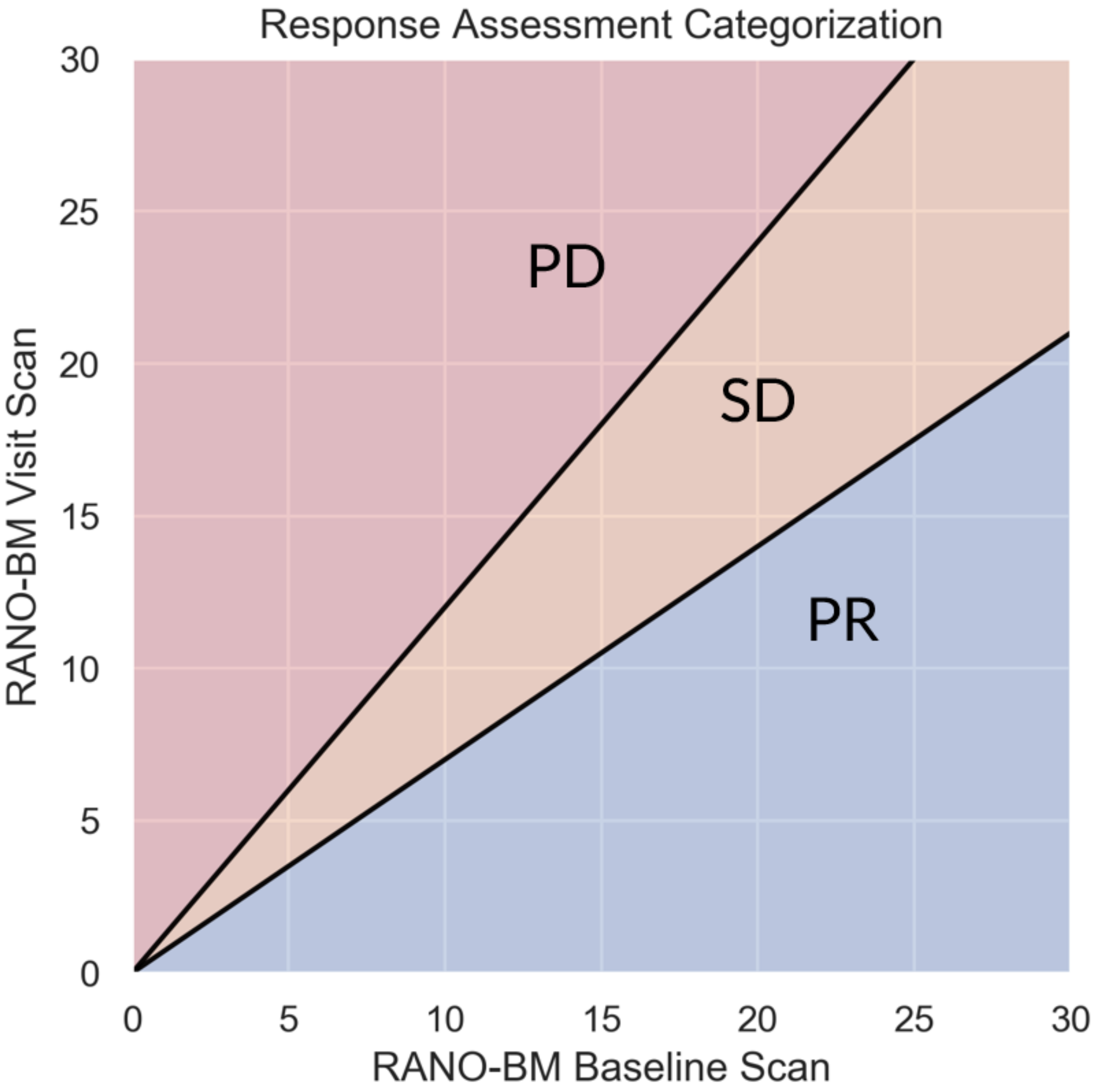


Figure 4-6: **Response assessment categorization based on RANO-BM.** Given a RANO-BM measure for the baseline and new timepoint scan, we can ascertain the response assessment.

not. Metastases that are present in prior scans are given the same identifier to enable individual lesion tracking across time. New metastases not present in prior scans are given an unused numeric identifier. Absolute change in volume per metastasis is then calculated. An example showing volumetric longitudinal tracking, along with AutoRANO-BM and response assessment categorization, is shown in figure 4-7. Briefly, we note subtle differences between the volumetric curves and the AutoRANO-BM curve for this example patient. The volumes curves change smoothly over time and depict that this patient is slowly progressing. Conversely, a sharp discontinuity is noted in the AutoRANO-BM curve due to the fact that lesions under 10 mm are not measured. In this specific case, the orange lesion is too small to be considered a target lesion as per the RANO-BM criteria up to the visit 2 scan. By the visit 3 scan, the lesion has now grown above the 10 mm threshold, allowing it to be measured and included in the sum total diameter measure. This discontinuity can lead to an alternative (and perhaps incorrect) interpretation that the patient is stable for most of the treatment course except for in between visits 2 and 3.

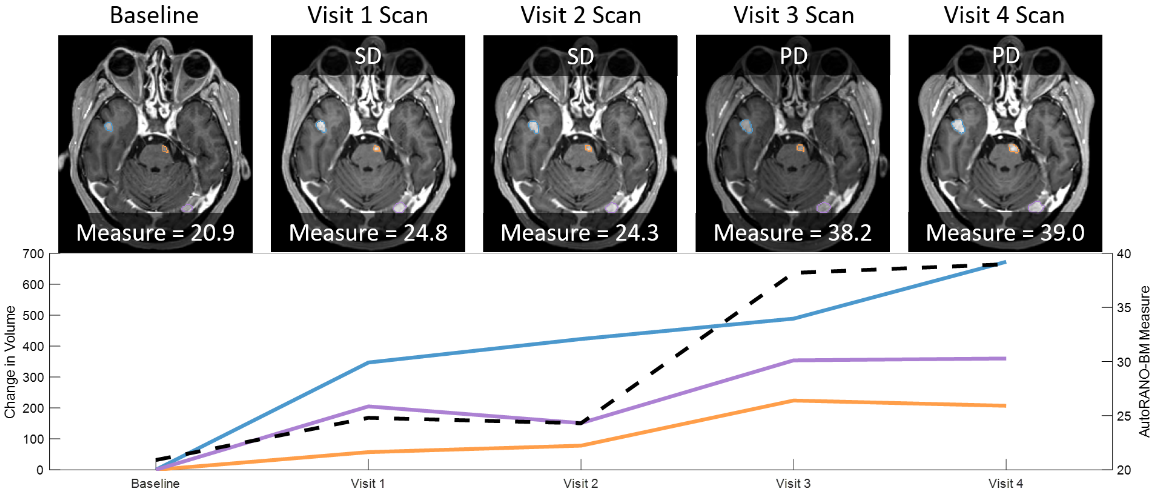


Figure 4-7: **Longitudinal tracking of metastases across time.** BM at each timepoint are automatically segmented. After co-registration to the baseline, each lesion is given a unique identifier to allow it to be tracked across timepoints. In this case, the three distinct lesions are color coded blue, orange, and purple. Solid colored lines show the absolute change in volume from baseline, showing that lesions can exhibit significantly different growth rates over time. The dashed black line shows the AutoRANO-BM measure for each timepoint. Response assessment categorization is shown above each timepoint image.

4.2.6 Uncertainty Estimation

We compute voxel-wise segmentation uncertainty [202] $U_v(x) \approx 0$ by taking the mean entropy over the output from each of the $N = 5$ models in our ensemble, implemented as follows:

$$U_v(x) = -\frac{1}{N}\sum_{i=1}^{N} p_i(x) \log_2 p_i(x) \tag{4.5}$$

$U_v(x)$ ranges from 0 to 0.5, where a voxel with high certainty will have $U_v(x) \approx 0$ and a voxel with high uncertainty will have $U_v(x) \approx 0.5$. Lesion-wise uncertainty $U_l(x)$ is computed by taking the median voxel-wise uncertainty over voxels in each unique tracked lesion.

4.2.7 Statistical Analysis

Neural network segmentation performance was evaluated using DSC and 95th percentile of Hausdorff distance (HD95). A Wilcoxon signed-rank test was used to evaluate significant pairwise differences in these metrics. Due to the difficulty in detecting and segmenting brain metastases, with specific regards to micro-metastatic lesions, we compute statistics on both the examination level (i.e. DSC over entire image) and the lesion level (i.e. DSC of each lesion in the image separately). Additionally, per lesion sensitivity was used to measure the detection rate of metastases. McNemar's test was used to compare lesion sensitivities. For longitudinal comparison of volume and AutoRANO-BM measures, both Pearson's correlation coefficient ρ and the two-way mixed, single measure intra-class correlation coefficient (ICC) were used [203]. To quantify the relationship between lesion size and detection classification (true positive (TP), false positive (FP), false negative (FN)), a two-way ANOVA with multiple pairwise comparisons using Tukey's HSD test was used. All statistical analyses were performed using python 3.6.9 [169].

4.3 Results

4.3.1 Deep Learning Based Segmentation

To assess the efficacy of our boundary-weighted cross-entropy loss, we compare segmentation performance when using different loss functions on the primary validation set, with examination level results and metastasis level results presented in tables 4.2 and 4.3, respectively. Specifically, we compare against dice loss, weighted cross entropy (with $\alpha = 0.25$), and focal loss [73] (with $\alpha = 0.25$ and $\gamma = 2$). Examination level DSC for boundary loss was significantly larger than that for cross entropy ($p < 0.01$) and focal loss ($p < 0.01$), but not for dice loss ($p = 0.09$). Examination level HD95 was not significantly smaller than that for any of the three other losses. Lesion level DSC was significantly larger and HD95 significantly smaller than that for all three other losses ($p < 0.001$). Sensitivity of lesion detection for boundary loss was significantly higher than that of all three other losses ($p < 0.01$).

Table 4.2: Comparison of examination level segmentation performance for different loss functions on the primary validation set of 32 cases.

Loss Functions	DSC†	HD95†
Dice Loss	0.81 (0.67 - 0.87)	1.73 (1.0 - 3.16)
Weighted Cross Entropy Loss	0.81 (0.68 - 0.86)	2.34 (1.41 - 4.79)
Focal Loss	0.80 (0.66 - 0.87)	2.24 (1.41 - 4.70)
Boundary Loss	0.82 (0.70 - 0.88)	1.87 (1.0 - 3.64)

† Data in parantheses are IQRs.

Full performance metrics for automatic segmentation of BM on the primary training, primary validation, primary testing, and secondary testing cohorts were computed at the examination level (table 4.4) and at the metastasis level (table 4.5). Additionally, metrics were computed at different metastasis size thresholds (table 4.6). An example showing manual and automatic segmentation is shown in figure 4-8. Importantly, manual segmentation is not always perfect, mainly because human annotators often make small mistakes. For instance, a human annotator might not perfectly outline the contrast enhancing portion of a tumor (perhaps due to fatigue), whereas a neural

Table 4.3: Comparison of metastasis level segmentation performance for different loss functions on the primary validation set of 379 metastases.

Loss Functions	DSC†	HD95†	Sensitivity*
Dice Loss	0.75 (0.52 - 0.85)	1.0 (1.0 - 3.0)	83 (79 - 87)
Weighted Cross Entropy Loss	0.77 (0.59 - 0.85)	1.0 (1.0 - 2.45)	84 (80 - 88)
Focal Loss	0.75 (0.56 - 0.84)	1.41 (1.0 - 2.45)	84 (80 - 87)
Boundary Loss	0.77 (0.63 - 0.86)	1.0 (1.0 - 2.24)	86 (82 - 89)

* Data in parantheses are percentages.

† Data in parantheses are IQRs.

network can consistently and accurately perform the same task repeatedly without variability in performance.

Table 4.4: Examination level segmentation performance on the four datasets.

Dataset	N	DSC†	HD95†
Primary Train	118	0.83 (0.78 - 0.87)	1.41 (1.0 - 3.0)
Primary Validation	32	0.81 (0.68 - 0.87)	1.73 (1.31 - 3.48)
Primary Test	32	0.80 (0.71 - 0.86)	1.57 (1.41 - 2.38)
Secondary Test	885	0.80 (0.70 - 0.86)	1.73 (1.41 - 11.41)

† Data in parantheses are IQRs.

Table 4.5: Metastasis level segmentation performance on the four datasets.

Dataset	N	DSC†	HD95†	Sensitivity*
Primary Train	1171	0.76 (0.45 - 0.84)	1.0 (1.0 - 4.0)	78 (76 - 81)
Primary Validation	379	0.78 (0.63 - 0.86)	1.0 (1.0 - 2.24)	84 (80 - 87)
Primary Test	548	0.76 (0.61 - 0.84)	1.0 (1.0 - 2.0)	86 (83 - 89)
Secondary Test	5250	0.64 (0.0 - 0.79)	1.41 (1.0 - inf)	73 (72 - 74)

* Data in parantheses are percentages.

† Data in parantheses are IQRs.

4.3.2 Longitudinal Tracking

To measure the quality of longitudinal lesion tracking, we compared ground truth and predicted volumes for all detected TP lesions in the secondary test set. Directly

Table 4.6: Metastasis level segmentation performance on the four datasets, split by volumetric size.

Dataset	Size	N	DSC†	HD95†	Sensitivity*
Primary Train	small	604	0.61 (0.0 - 0.75)	1.41 (1.0 - inf)	61 (57 - 65)
	medium	385	0.81 (0.75 - 0.86)	1.0 (1.0 - 1.41)	96 (93 - 97)
	large	182	0.88 (0.82 - 0.92)	1.0 (1.0 - 2.0)	98 (94 - 99)
Primary Validation	small	154	0.62 (0.0 - 0.76)	1.41 (1.0 - inf)	62 (54 - 70)
	medium	172	0.83 (0.77 - 0.87)	1.0 (1.0 - 1.41)	98 (95 - 99)
	large	53	0.89 (0.85 - 0.92)	1.0 (1.0 - 1.73)	100 (93 - 100)
Primary Test	small	258	0.64 (0.0 - 0.74)	1.38 (1.0 - inf)	71 (66 - 76)
	medium	247	0.81 (0.75 - 0.87)	1.0 (1.0 - 1.41)	99 (96 - 100)
	large	43	0.88 (0.81 - 0.93)	1.41 (1.0 - 1.57)	100 (92 - 100)
Secondary Test	small	3161	0.45 (0.0 - 0.66)	2.0 (1.06 - inf)	59 (58 - 61)
	medium	1367	0.78 (0.70 - 0.83)	1.25 (1.0 - 1.73)	92 (90 - 93)
	large	722	0.87 (0.83 - 0.90)	1.41 (1.0 - 2.24)	98 (96 - 98)

* Data in parantheses are percentages.

† Data in parantheses are IQRs.

comparing between ground truth and predicted lesion volume, the ICC was 0.92 (95% CI: 0.91 – 0.92) and the Pearson correlation ρ was 0.92. Comparing the absolute change in lesion volume across consecutive patient timepoints between ground truth and predicted segmentations, the ICC was 0.88 (95% CI: 0.87 – 0.88) and the Pearson correlation ρ was 0.88 (figure 4-9).

4.3.3 RANO-BM

In assessing agreement between manual and automatic measurements for total tumor volumetric burden, the ICC was 0.91 (95% CI: 0.89 – 0.92) and the Pearson correlation ρ was 0.91. When assessing agreement for target tumor volume, the ICC was 0.92 (95% CI: 0.91 – 0.93) and the Pearson correlation ρ was 0.92. Direct comparison between AutoRANO-BM measures computed from ground truth segmentations and from predicted segmentations yielded an ICC of 0.92 (95% CI: 0.91 – 0.93) and the Pearson correlation ρ was 0.92 (figure 4-10). Example segmentations and AutoRANO-BM diameter measurements shown in figure 4-11.

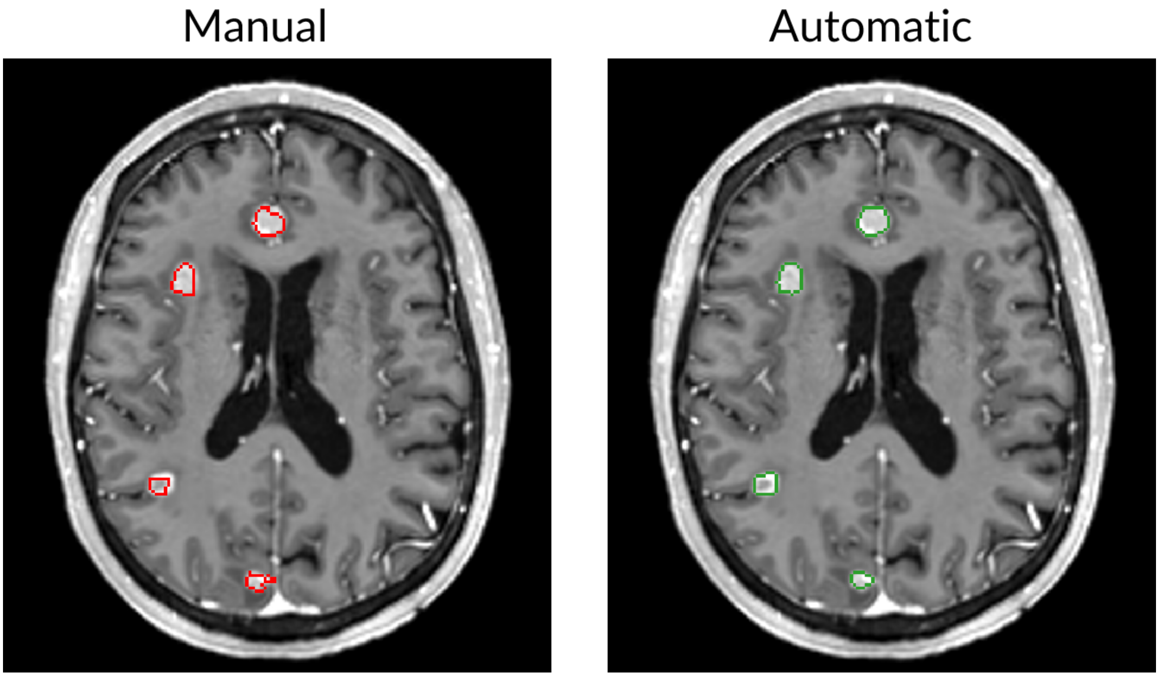


Figure 4-8: **Manual vs automatic segmentation of brain metastases patient.** Our automatic segmentation method does make mistakes, and can often miss small micro-metastases. However, we note that automatic segmentation can in fact be more accurate than manual segmentation in some cases, as can be seen here where the manual segmentation is not completely encircling the contrast enhancing portion of the tumor.

4.3.4 Response Assessment Classification

Response assessment classification based on AutoRANO-BM measures was computed and agreement is shown in the confusion matrix in figure 4-12. Response assessment was computed for all timepoints excluding the baseline, resulting in 737 data points. Forty-eight of 67 (72%) were correctly categorized as PR, 72 of 112 (64%) were correctly categorized as SD, and 506 of 558 (91%) were correctly categorized as PD. No examinations had a true or predicted class of CR.

4.3.5 Uncertainty Estimation

Lesion-wise uncertainty was calculated and then stratified by lesion size and detection classification (shown in figure 4-13). We conclude that lesion size ($p < 0.001$), detection

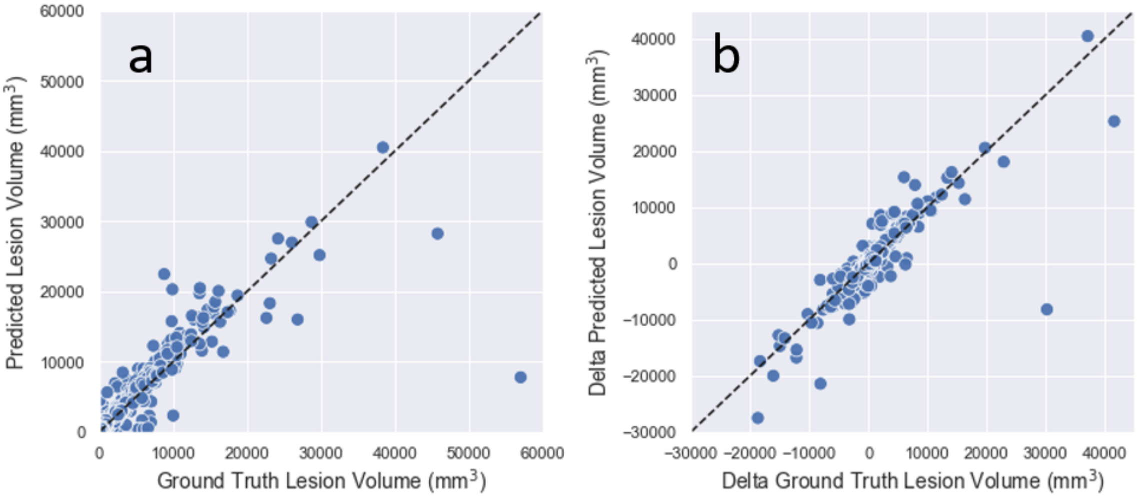


Figure 4-9: **Volumetric comparison between ground truth and predicted segmentations.** (a) We plot ground truth vs. predicted metastases volume. (b) The absolute change in ground truth lesion volume across consecutive timepoints was calculated and plotted against the change in predicted lesion volume. Line of identity (x = y) is shown in all plots.

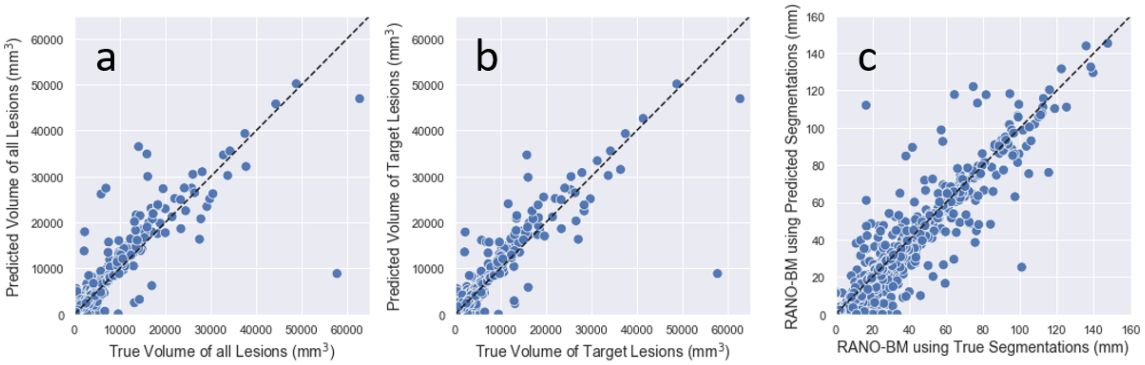


Figure 4-10: **Comparison between AutoRANO measurements from ground truth and predicted segmentations.** (a) Scatter plot showing total metastases volume for ground truth and automatic segmentations. (b) Scatter plot showing target metastases volume for ground truth and automatic segmentations. (c) Scatter plot showing AutoRANO-BM measure for ground truth and automatic segmentations. Line of identity (x = y) is shown in all plots.

classification ($p < 0.001$), and the interaction between the two ($p < 0.001$) significantly affects lesion-wise uncertainty. Moreover, we note that while there is a significant difference in the uncertainty between FP and TP lesions (($p < 0.001$)) and FP and TN

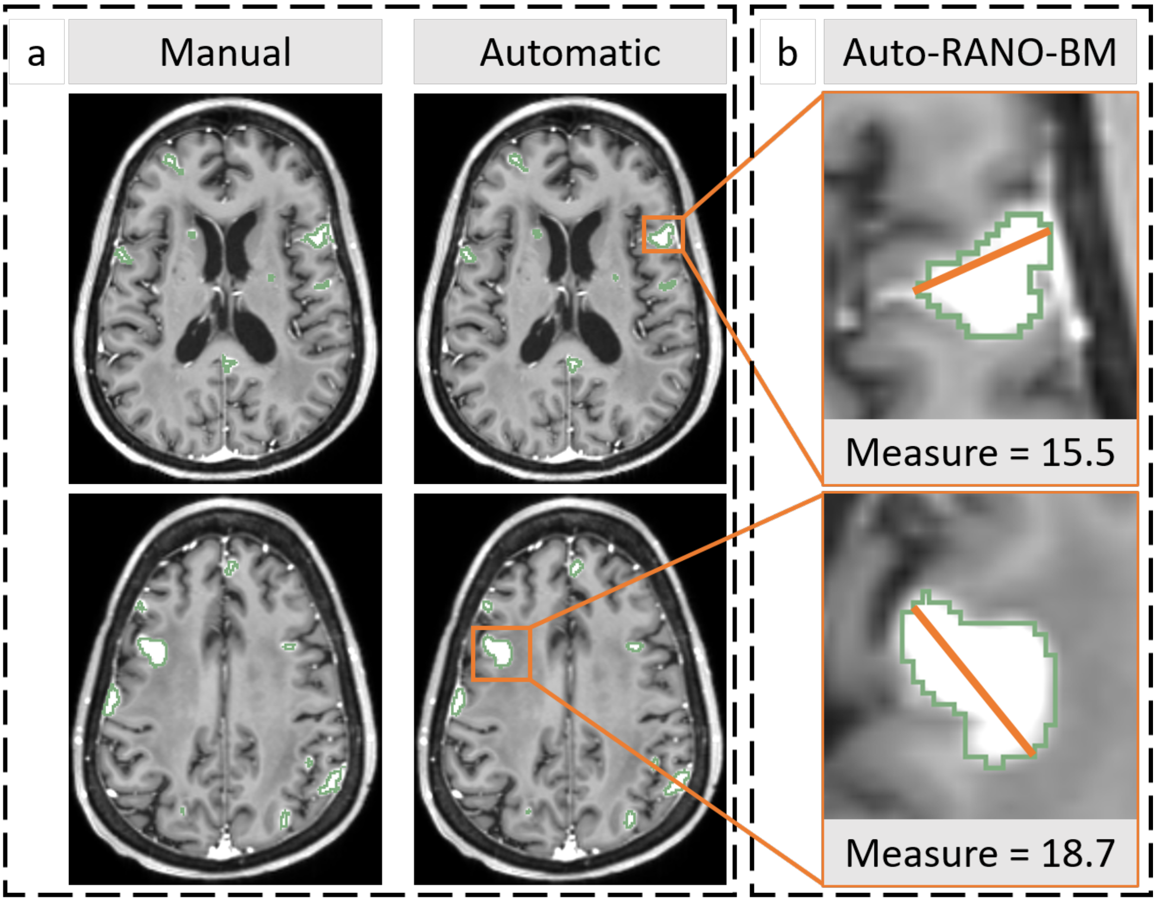
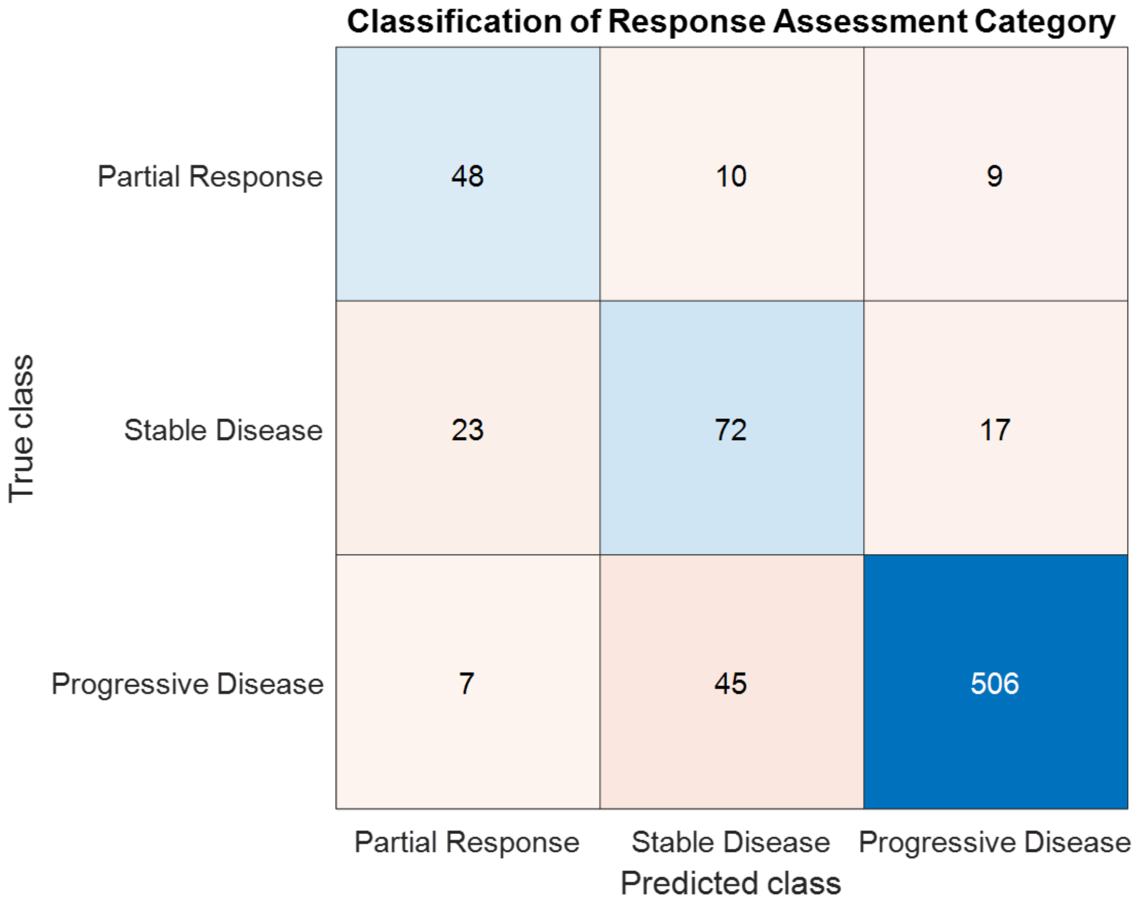


Figure 4-11: **Examples of automatic segmentation and AutoRANO-BM.** (a) Examples of manual vs automatic segmentation of BM. (b) Examples of an automatically segmented lesion with the automatically calculated longest uni-dimensional diameter drawn through it.

lesions ($p < 0.001$), we fail to find a significant difference between TP and FN lesions ($p = 0.51$). An example of a low uncertainty and a high uncertainty lesion is shown in figure 4-14. In the second column of this figure, our segmentation model accidentally segments a non-tumor abnormality (most likely a central pontine myelinolysis). We note that our lesion-wise uncertainty model can automatically flag this case as a potential false positive.



	Partial Response	Stable Disease	Progressive Disease
Partial Response	48	10	9
Stable Disease	23	72	17
Progressive Disease	7	45	506

Figure 4-12: **Confusion matrix for response assessment classification.** Classification of Partial Response (PR), Stable Disease (SD), and Progressive Disease (PD) using AutoRANO-BM measures from the ground truth segmentations and predicted segmentations.

4.4 Discussion

Given the difficulty in manually annotating brain metastases, the goal of this study was to create a tool that could aid clinicians by providing high quality 3D segmentations for use in calculation of individual tumor volumes across time and automated RANO-BM measures. To demonstrate model performance, we evaluated our model on a large and diverse secondary dataset obtained retrospectively for adult patients with newly diagnosed brain metastases. This set contained examinations which ranged from having anywhere from only 1 to up to 445 unique lesions and contained lesions which varied in volume from as small as a few mm^3 to greater than 50000mm^3. Moreover,

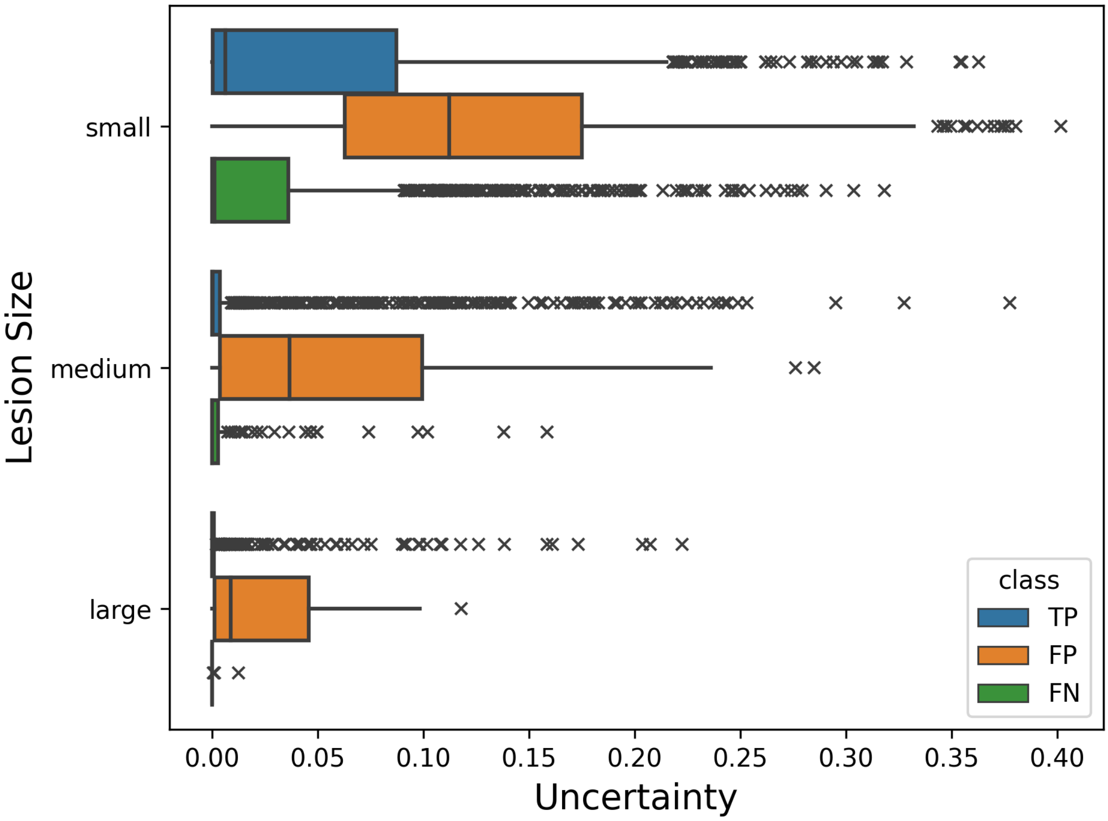


Figure 4-13: **Lesion-wise uncertainty split by lesion size and detection classification.** We note that FP lesions exhibit significantly higher uncertainty than either TP or FN lesions.

lesions exhibited varied shapes/structures (from spherical to highly irregular), were situated across every region of the brain parenchyma, and originated from several types of primary cancers. Thus, this dataset spans the variability in number, size, shape, and location of brain metastases seen in clinical practice, and as such provides an excellent dataset by which to robustly judge the clinical utility of our model.

We tested a few common segmentation loss functions and evaluated their performance at both the examination level and lesion level. We note that boundary-weighted cross-entropy significantly improves sensitivity of lesion detection, indicating that it is better suited to identifying and segmenting hard to detect lesions compared to other standard loss functions such as dice, weighted cross-entropy, or focal loss.

While we note that the segmentation model was highly performant for medium and large sized lesions as evidenced by high sensitivity and DSC, its performance suffered for small lesions. This drop in performance is attributed to a couple of factors. First, brain micro-metastases (especially dural lesions located at the peripheries of the brain) can share similarity in shape, size, and MR intensity to small blood vessels. In such cases, it can be challenging to confidently label a small focus of enhancement as a metastasis until it grows on a subsequent timepoint. In other cases, micro-metastases can present with little to no contrast enhancement, especially in the setting of lower quality scans, and can be missed entirely on lower resolution scans. This annotation challenge can lead to missed micro-metastases in the ground truth, which can negatively affect the training of the neural network. While the detection rate for micro-metastases is low, we note that these small lesions are potentially of less clinical importance. As per the RANO-BM guidelines, a target lesion is defined as any contrast enhancing lesion with at least a 10mm diameter visible on two or more axial slices. Thus, smaller lesions are generally not tracked for the purposes of response assessment. Moreover, since micro-metastases usually represent a small fraction of a patient's full volumetric tumor burden, patient treatment plans may not change due to some missed detections. For instance, our model detected 26% (39 of 149) unique lesions for a patient examination, and these detected lesions accounted for 72% (1494mm^3 of 2074mm^3) of the full tumor burden. Indeed, the treatment for a patient with 39 lesions or 149 lesions is the same - whole brain radiation.

Our model is capable of not only tracking total tumor volumetric burden across time, but also individual lesion volume across time. High correlation was seen when comparing absolute volumetric change per lesion across time for manual and automatic segmentations. This indicates that we can track the growth rates of individual brain metastases over time with sufficiently reasonable accuracy.

There was high agreement between manual and automatic measures with regard to changes in tumor burden (AutoRANO-BM, target tumor volume, and total tumor volume). While all three objectively exhibit almost the same ICC and Pearson correlation values, we note that subjectively, RANO-BM measures are clustered less tightly along the identity line than are the two volumetric measures. The reason for the unexpectedly similar correlation is due to the effect of two outlier lesions in our dataset. The first was a large cystic lesion (the biggest individual lesion in our dataset) and the second was a highly heterogeneous tumor with complicated and blurred boundaries. Due to not having any training data examples similar to these two rare cases, our model was not capable of segmenting them properly. While these types of large cystic lesions are clinically meaningful treatment targets, as per the RANO-BM criteria, they are considered not measurable for clinical trial purposes due to the extensive necrotic regions present. Nevertheless, the overall trend suggests that when comparing manual and automatic measures, volumetric analysis allows for greater agreement than the RANO-BM measure.

There was high accuracy for the classification of each patient timepoint as non-progressive disease (CR, PR, and SD) and progressive disease (PD) according to the RANO-BM criteria using our automatically computed RANO-BM measures. This further serves to indicate that the relatively low detection rate of micro-metastases does not significantly affect radiographic response assessment.

Due to the fact that RANO-BM is the sum of up to five uni-dimensional diameter measurements, it can be an imperfect proxy for true volumetric tumor burden. An example with four patients, each with a very similar total volumetric tumor burden, is shown in figure 4-15. Due to the fact that these patients have a different number of metastases each, ranging from 1 to 15, the RANO-BM measures are completely different from each other. The patient with 15 micro-metastases has a RANO-BM measure of 0 even though the total tumor burden is similar to the other patients. Consideration should be made about whether RANO-BM is truly the best metric on which to base treatment response assessment, seeing that it is a potentially incomplete view tumor burden.

One concern of deep learning based models when used in clinical care settings is that they tend to be un-explainable and can often generate outputs that are overconfident of incorrect results. To mitigate this issue, we study the lesion-wise segmentation uncertainty of our model and show that predicted lesions that are actually FPs exhibit higher median uncertainty than do TPs and FNs. By producing a lesion-wise uncertainty map of the image, this may serve as an automatic way of flagging specific lesions for manual expert review.

There have been several prior studies investigating the ability of deep learning algorithms to automatically segment metastatic brain tumors. Grovik et al. [140] trained a modified GoogLeNet architecture, which took four distinct MR sequences as input, on a single-center set of 156 patients. This study reported a sensitivity of 50% of lesions smaller than 7 mm. Ottesen et al. [189] expanded on the prior study from Grovik et al. [140] by utilizing a secondary dataset of 65 patients in addition to the primary set of 156 patients. They test both 2.5D and fully 3D approaches and investigate model detection and false positive rates. While they do not share quantitative results split by lesion size, they do note reduced segmentation performance for smaller metastases, especially in their secondary test set. Rudie et al. [190] ensembled models using different inputs and different loss functions together to boost overall performance. On a test set of 100 patients, they show an overall sensitivity of 70%, with a sensitivity of 50.9% for lesions less than 5mm. While direct comparison between our work and previous studies is confounded by factors such as the relative proportions of small, medium, and large lesions changing between different private datasets, our model achieves segmentation performance equivalent or better than that of networks in recently published literature. Moreover, our independent secondary testing set is one of the largest datasets used for this task, and we go beyond simply reporting model detection rates by proposing a method to capture lesion-wise uncertainty to automatically identify FPs.

4.5 Limitations

Our study has limitations. First, our primary dataset was segmented by a single rater, which limits our ability to assess interrater variability and may cause the neural network to learn certain biases specific to that rater. Moreover, the training set was comprised of single-center data acquired under strict criteria for clinical trials. Future work could entail utilizing a larger, multi-institutional cohort for training to improve model generalizability and robustness. Secondly, our model was trained to segment contrast enhancing disease and some BM can be cystic or hemorrhagic without clear additional enhancement. These cases would not be captured with our model but could be explored in future model optimizations. Finally, we did not have any manual RANO-BM measures to compare our AutoRANO-BM measures against. Chang et al. [187] did an in depth analysis of RANO for primary gliomas and concluded that AutoRANO was significantly more reliable and consistent then were manual RANO measurements calculated by expert neuro-oncologists and that it was more correlated to total tumor volumetric burden. Peng et al. [204] noted a similar trend when comparing manual and automatic measurements for pediatric gliomas. Still, future studies could explore the correlation between manual and automatic RANO measurements in the setting of brain metastases.

4.6 Conclusion

In conclusion, we developed a fully automatic pipeline that segments BM, computes lesion uncertainty, longitudinally tracks individual lesions across time, and computes an automatic RANO-BM measure. Moreover, we show high agreement between the AutoRANO-BM measure computed from the ground truth segmentations and the predicted segmentations, demonstrating the clinical utility of our automated method to quantify measurable tumor burden.

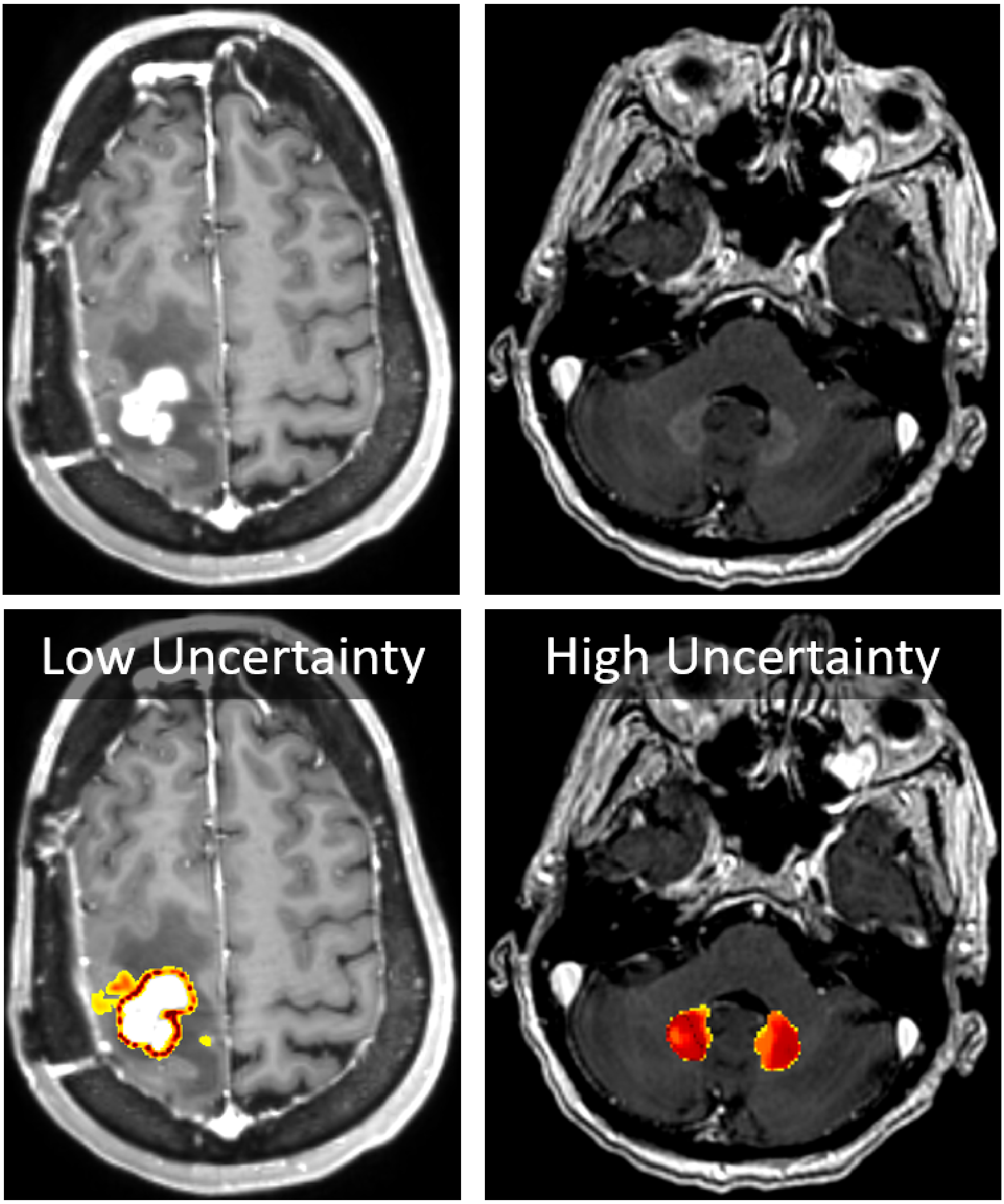


Figure 4-14: **Low vs high uncertainty example images.** Uncertainty is represented by color, with yellow indicting low uncertainty and red indicating high uncertainty. The first column shows an examination with a clear metastasis, which the model correctly identifies and segments with little uncertainty. The second column shows an examination with a non-tumor abnormality. Our model accidentally segments this abnormality, but is highly uncertain of its prediction.

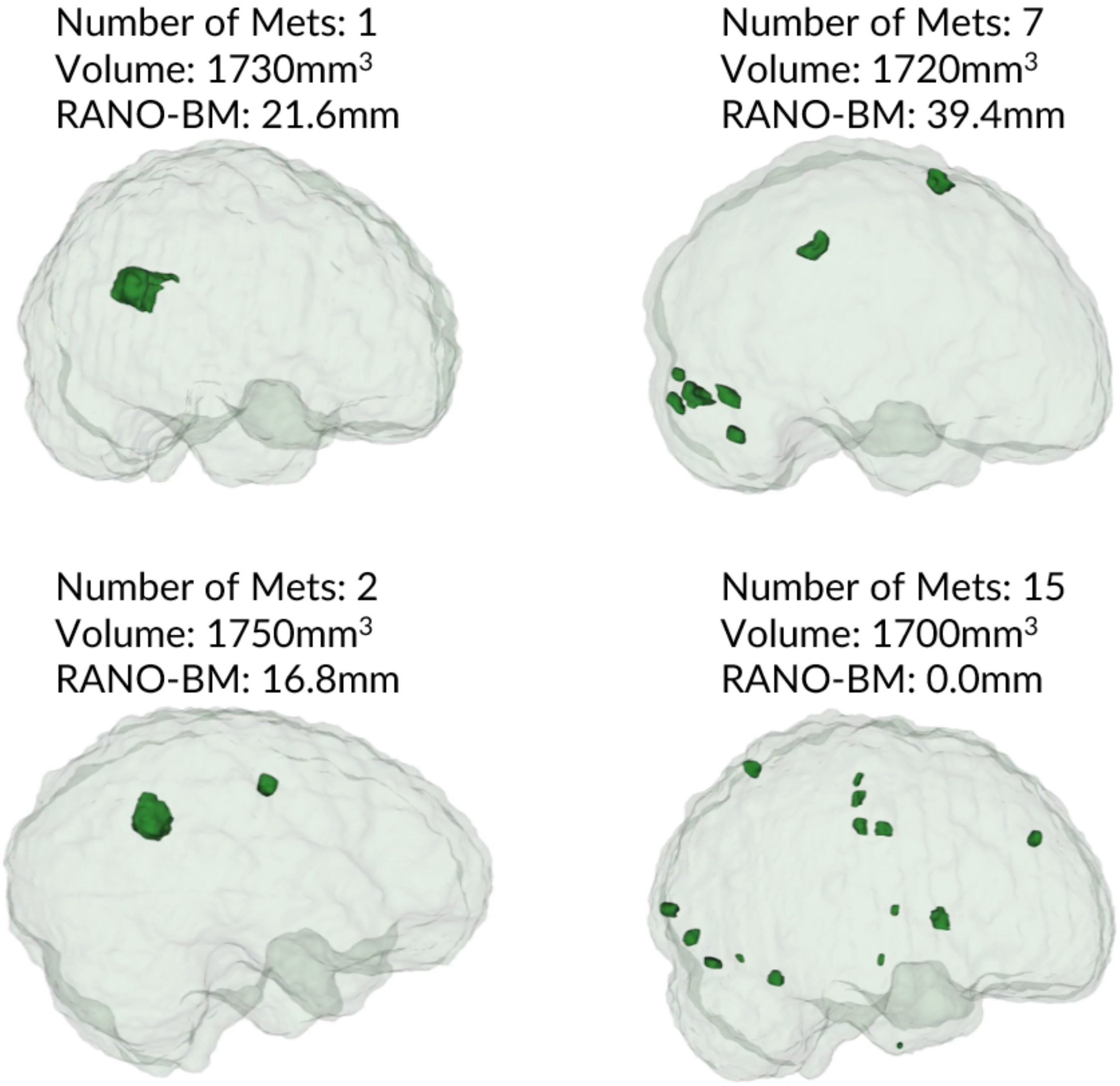


Figure 4-15: **Volume vs RANO-BM measurements.** RANO-BM is a potentially incomplete view tumor burden, seeing that patients with similar total tumors volume can have drastically different RANO-BM measurements.

Chapter 5

Joint Image Registration and Segmentation

Manual segmentation of medical imaging is a laborious and time-consuming task for expert clinicians, especially in the setting of longitudinal patient data. Deep learning (DL) based segmentation approaches are becoming increasingly common as researchers push the boundaries of artificial intelligence (AI) systems in healthcare applications. Indeed, many algorithms now boast human or super-human performance on a wide range of segmentation tasks [205, 206, 207, 208]. However, these approaches tend to segment all patient imaging independently of each other, ignoring relevant information from prior time-points. In order to utilize prior time-point information, we propose a joint image registration and segmentation method. Given a prior time-point image and segmentation mask (which are readily available in a routine clinical environment), we affinely and deformably register these to a new time-point image. This warped prior image and mask are then used to enhance and improve the segmentation of the new time-point. In this chapter, we develop a novel deep learning based framework for joint image registration and segmentation and apply it to brain metastases segmentation, aiming to improve segmentation performance for micro-metastatic lesions.

5.1 Introduction

Brain metastases (BM) are the most common form of intracranial tumors in adults with an annual incidence of 170,000 in the United States [209], which is expected to increase as systemic treatments for primary tumors improve [177, 210]. Given rising incidence and limited treatment, BM are an unmet need in modern oncology, with median survival post diagnosis of ranging from only 2.7 to 24 months [211, 212]. Alongside monitoring changes in clinical metrics such as performance status and cognitive function, neuroradiologists can assess the efficacy of a given treatment regimen by tracking individual lesion sizes across T1-weighted contrast-enhanced (T1-CE) magnetic resonance (MR) imaging time-points, noting whether tumor burden is decreasing, stable, or increasing [213]. If BM enlarge or new BM appear over time, a different treatment option may be necessary. However, manual determination of tumor boundaries needed for lesion tracking can be challenging in the presence of heterogeneous contrast enhancement, diffuse tumor boundaries from surrounding edema, or blunted contrast relative to surrounding normal brain due to treatment effects. Moreover, patients can present with anywhere from a single lesion to upwards of one hundred lesions, varying in volume from as small as a few mm^3 to as large as 10000mm^3. These lesions can exhibit varied shapes/structures (from spherical to highly irregular) and can be situated across every region of the brain parenchyma. In addition to being a highly time-consuming and costly task, manual segmentation is subject to significant inter- and intra-rater variability for the aforementioned factors [180]. As such, there has been much interest in developing reproducible automated methods for segmentation.

While there is minimal work published in the automated segmentation of *metastatic* brain tumors, the segmentation of *primary* brain tumors is a well-researched field, mainly due to the availability of the large multi-institutional publicly available BraTS dataset [94, 214, 215]. In recent years, 3D U-Net architectures [99] have consistently dominated the BraTS leaderboards and are the current state-of-the-art method for brain tumor segmentation [100, 101, 102]. Guided by these approaches, most researchers also choose to use 3D U-Nets for BM segmentation. Due to the fact that BM can vary in size and number of lesions, *metastatic* brain tumor segmentation is a more difficult task than *primary* brain tumor segmentation. More specifically, the difficulty is attributed to the presence of micro-metastases, which are lesions with a diameter no greater than 5mm [216, 217]. These brain micro-metastases (especially dural lesions located at the peripheries of the brain) can share similarity in shape, size, and MR intensity to small blood vessels. In such cases, it can be challenging to confidently label a small focus of enhancement as a metastases or a blood vessel until it grows on a subsequent time-point. In other cases, micro-metastases can present with little to no contrast enhancement, especially in the setting of lower quality scans. These factors together make the automated detection and segmentation of micro-metastatic brain lesions a challenging machine learning problem.

Past approaches in published literature on the automated segmentation of BM have included a variety of architectural modifications and different loss functions, but they all report a sensitivity of detection of micro-metastases well under 50% [218, 189, 190]. Studies that have looked at the inter-rater variability for detection of micro-metastases have concluded that current deep learning (DL) based approaches are inferior to that of expert neuroradiologists and might not be ready for clinical deployment just yet [190]. That being said, even though current models are imperfect, especially for detection of small lesions, they have the potential to improve workflow efficiency for radiologists by reducing the amount of manual segmentation that must be done by a clinician. Specifically, a clinician can now simply correct a label map by adding in missed detections or removing false positives, a process that can save significant amounts of time relative to needing to segment the whole volume from scratch.

While the process of correcting mistakes is acceptable in the setting of single patient visits, it can become cumbersome and tedious to the clinician in the setting of longitudinal patient data. For instance, a metastasis that is missed in the baseline scan is likely to be missed again in all future time-points assuming it does not substantially change in size or appearance. Such a scenario requires the clinician to manually fix the same mistake repeatedly on all time-points, creating unnecessary annotation burden for the clinician. A better system would entail the clinician fixing the mistake only once on the first time-point, with the neural network carrying forward that prior information to subsequent time-points. An schematic of this concept is shown in figure 5-1. To the best of our knowledge, no published work has assessed the utility of using longitudinal imaging data for the purpose of improving BM segmentation quality.

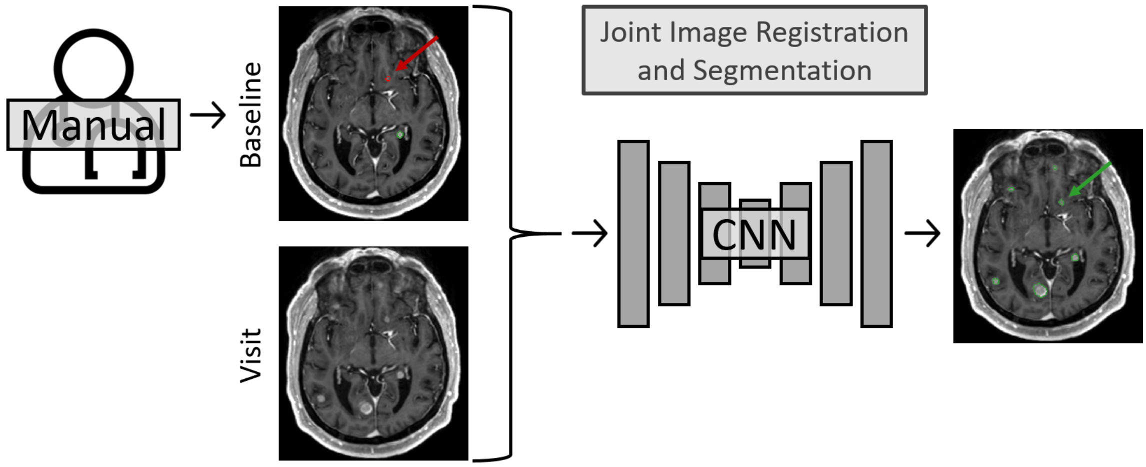


Figure 5-1: **System where clinician only needs to fix the mistake once.** In this schematic, the clinician manually corrects the baseline scan, adding in a missed micro-metastasis in red. The next timepoint for this patient is then jointly segmented with this prior corrected label mask to ensure that the model does not miss the micro-metastasis again on the new timepoint.

In this work, we propose a novel DL based approach to jointly register and segment BM on T1-CE MR imaging, a method we call Sequential and Pyramidal Image Registration and Segmentation (SPIRS). More specifically, given a prior time-point and a new time-point image, we train a *Siamese* style convolutional neural network (CNN) to first affinely (i.e. linearly) and then deformably (i.e. non-linearly) register the pair of images. This registration transform is parameterized as a dense displacement vector field (DVF), and it maps the offset from the prior time-point onto the new time-point image. Assuming we already have a prior time-point segmentation mask that has been manually edited by a clinician (which will be the case in a routine clinical environment), we can then use the found DVF to transform this prior segmentation mask onto the new time-point. This warped prior mask can then be used to enhance and improve the segmentation of the new time-point (figure 5-2).

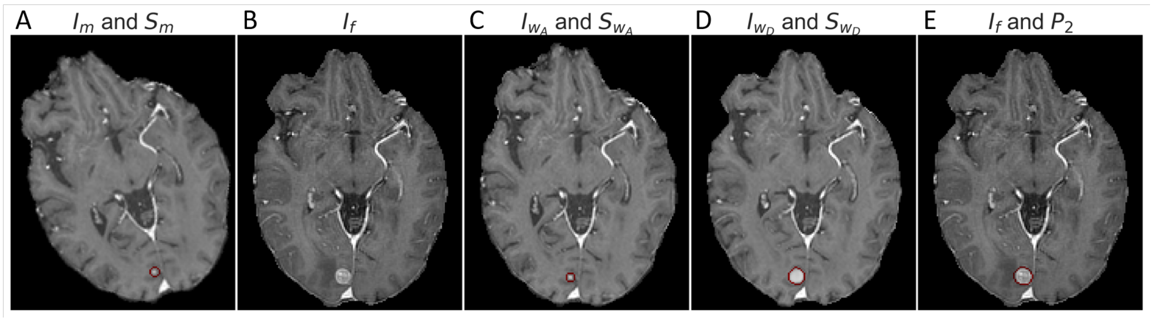


Figure 5-2: **Example outputs of our joint registration and segmentation model.** An example prior time-point image I_m and tumor mask S_m outlined in red (A) and new time-point image I_f (without tumor mask) (B). The priors are first affinely registered (creating I_{w_A} and S_{w_A}) (C) and then deformably registered (creating I_{w_D} and S_{w_D}) (D) to the new time-point. Note how the prior time-point tumor is warped to match the location and size of the new time-point tumor. These warped priors I_{w_D} and S_{w_D} are used to aid and improve the segmentation of the new time-point (E).

5.2 Generalizable Insights about Machine Learning in the Context of Healthcare

- **We show that SPIRS outperforms other methods for the specific task of registration of MR imaging with BM.** Many registration methods exist, but they are trained and validated on normal brain anatomy. Not only do we develop a novel architecture for combined affine and deformable registration, we show that by training a task specific model, we can improve performance over current baseline methods.

- **Our approach utilizes readily available prior time-point information in order to improve the segmentation of micro-metastases.** Despite the fact that longitudinal imaging data is readily available, most approaches segment scans independently of each other. Our method shows that using prior time-point information can significantly improve the detection rate of micro-metastases on follow-up imaging, which will reduce annotation burden for clinicians and improve performance of downstream clinical tasks.

- **Our model architecture is generalizable to other clinical applications with longitudinal data.** While we focus on BM for this manuscript, we emphasize that this approach can be applied to other challenging medical imaging segmentation problems where longitudinal imaging data is present. For instance, other tumor applications include brain meningiomas and other non-tumor applications include cardiac segmentation for congenital heart disease patients.

5.3 Related Work

There is extensive literature in medical image registration and it can be broadly be split into two categories: non-learning based and learning-based. An in-depth overview is provided in section 2.4. An abridged overview of relevant methodology is provided below.

5.3.1 Image Registration Approaches

Non-Learning Based Registration

Given a fixed and a moving image, classical registration approaches perform a gradient descent based numerical optimization to iteratively align pixels from the moving image onto the fixed image to improve a chosen similarity metric (e.g. mean squared error (MSE), normalized cross correlation (NCC)). The learned transformation can be either linear or non-linear, depending on one's use case. If the transform is non-linear, certain constraints can be placed on the outputted DVF to encourage a spatially smooth transform. To alleviate this numerical optimization problem (which even for linear transforms can get stuck at poor local minimas for anatomically complex images), classical methods often employ a *sequential* and *pyramidal* hierarchy. *Sequential* refers to solving lower complexity transforms before higher complexity transforms. In other words, a purely affine transformation is computed first before solving for the deformable transformation. *Pyramidal* refers to a multi-scale approach wherein the transformation is first computed at a coarser image scale and is progressively updated at finer image scales [114]. An example image pyramids are shown in figure 5-3. We note that due to the iterative nature of these classical algorithms, they can be quite computationally intensive. Indeed, deformable registration of 3D brain MR imaging can take upwards of one to two hours on CPU per image pair. While there are many classical registration algorithms for deformable registration, including but not limited to B-splines [115], Demons [116], and Large Diffeomorphic Distance Metric Mapping (LDDMM) [117], the current gold standard is generally accepted to be Symmetric Normalization (SyN) [118] from the Advanced Normalization Tools (ANTs) package [119, 120]. ANTs can also be used for highly performant affine registration of imaging.

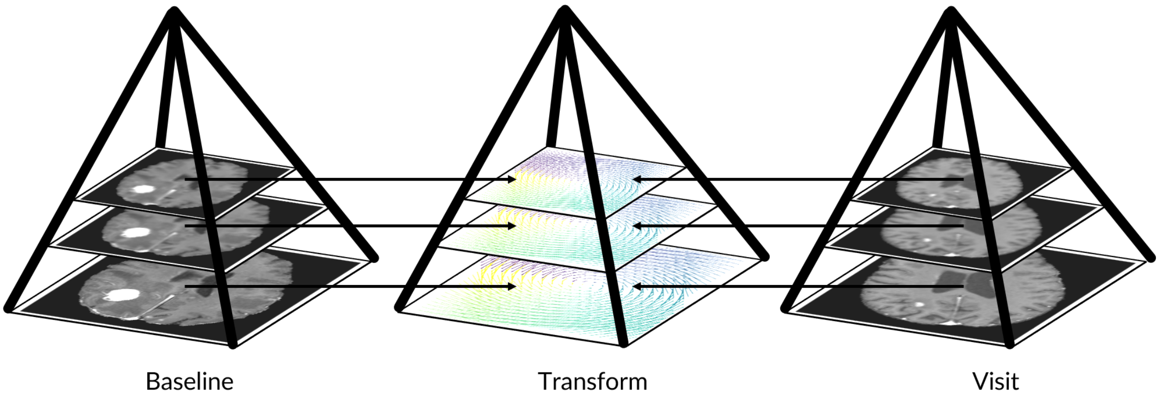


Figure 5-3: **Image Pyramids.** Image pyramids for the baseline and visit scan are created by stacking images with different resolutions and sizes. The registration is first run on the first level of the pyramid (the coarsest scale), and then iteratively updated at finer scales.

Learning Based Registration

Newer methods utilize neural networks to learn a function for (affine and/or deformable) registration. This can be advantageous because each image pair can be fully registered with one forward pass of the network, which will take only a few seconds on GPU. DL based affine registration networks are usually formulated as a supervised regression problem. [121] use a *Siamese* style encoder to directly predict the affine transform matrix. [122] uses a similar regression approach, but focuses on cross-modality registration. DL based deformable registration networks can either be trained in a supervised or unsupervised manner. While earlier approaches like that from [123] required ground truth DVFs to train the network, newer approaches tend to be fully unsupervised. [124] proposed a U-Net based diffeomorphic registration model they named VoxelMorph (VXM). Building off this approach, [125] utilized a pyramidal architecture to improve the quality of the registration. However, they did not incorporate feature sharing at the different levels of the pyramid, resulting in redundant parameters. [126] devised a network to sequentially perform affine and deformable registration, but their deformable registration was based only on b-spline grids. [127] also performed both affine and deformable registration, however their approach was neither sequential nor pyramidal.

5.3.2 Joint Frameworks for Registration and Segmentation

While registration and segmentation are two of the largest and most researched areas of computer vision for medical applications, there is significantly less research in how the coupling of these two tasks may improve one or both tasks. Such joint methods are mainly used in areas where longitudinal imaging data is widely available. For instance, segmentation of cardiac MR imaging is usually done only on end-diastolic and end-systolic frames, with information from other frames not exploited. By using a joint framework, more data may be incorporated during model training and can improve performance of both cardiac motion estimation and atrial/ventricular segmentation [219, 220]. Joint approaches have also been shown to be effective in low-annotation settings, where only a fraction of the whole dataset has ground truth segmentations [221]. When compared to the baseline of not having any annotated data, weak supervision from a small sample of ground truth annotations can improve registration performance due to the incorporation of an anatomy similarity loss. Indeed, joint methodological approaches have been shown to outperform independently optimized networks on tasks such as cardiac, knee, and brain [221, 222]. While these existing works have applied their respective methods to normal anatomy, to the best of our knowledge no work has focused specifically on improving the registration quality and segmentation performance of BM on T1-CE MR.

5.4 Methods

We let I_f, I_m, S_f, S_m denote the fixed image, moving image, fixed segmentation, and moving segmentation, respectively. $\hat{T}_A$ and $\hat{T}_D$ denote the predicted affine and deformable registration transforms, respectively. Here, $\hat{T}_A$ and $\hat{T}_D$ are mappings such that $I_m \circ \hat{T} = I_w \approx I_f$. To warp image I_m with transformation $\hat{T}$, we use a fully differentiable spatial transformer module, which allows for gradient backpropagation during network optimization [223]. Our proposed architecture consists of three successive blocks: 1) the affine registration network $\mathcal{F}_A$, 2) the deformable registration network $\mathcal{F}_D$, and 3) the segmentation network $\mathcal{F}_S$. A diagram showing this sequential structure is visualized in figure 5-4. We note that these three blocks all share the joint *Siamese* style feature encoder, which helps prevent overfitting towards any one task, since the model must learn encoded feature representations that are useful for all three tasks. We describe in detail the full CNN architecture and the training optimization strategies in the following sections.

5.4.1 Shared Encoder Architecture

In lieu of training three separate task specific encoders for each of the sub-networks, we instead train a single encoder which is shared between the tasks. This encoder is composed of 5 blocks, where each block consists of a batch normalization operation [164], ReLU activation [163], and kernel size 3 convolution with a stride of 1. To ensure our encoder learns robust yet powerful representations of the input data, we use 64 filters for the convolution in the first block, and double this number of filters as we go deeper into the network. Feature map downsampling occurs after each block and is accomplished through a max pooling operation [166] with a kernel size and stride both equal to 2.

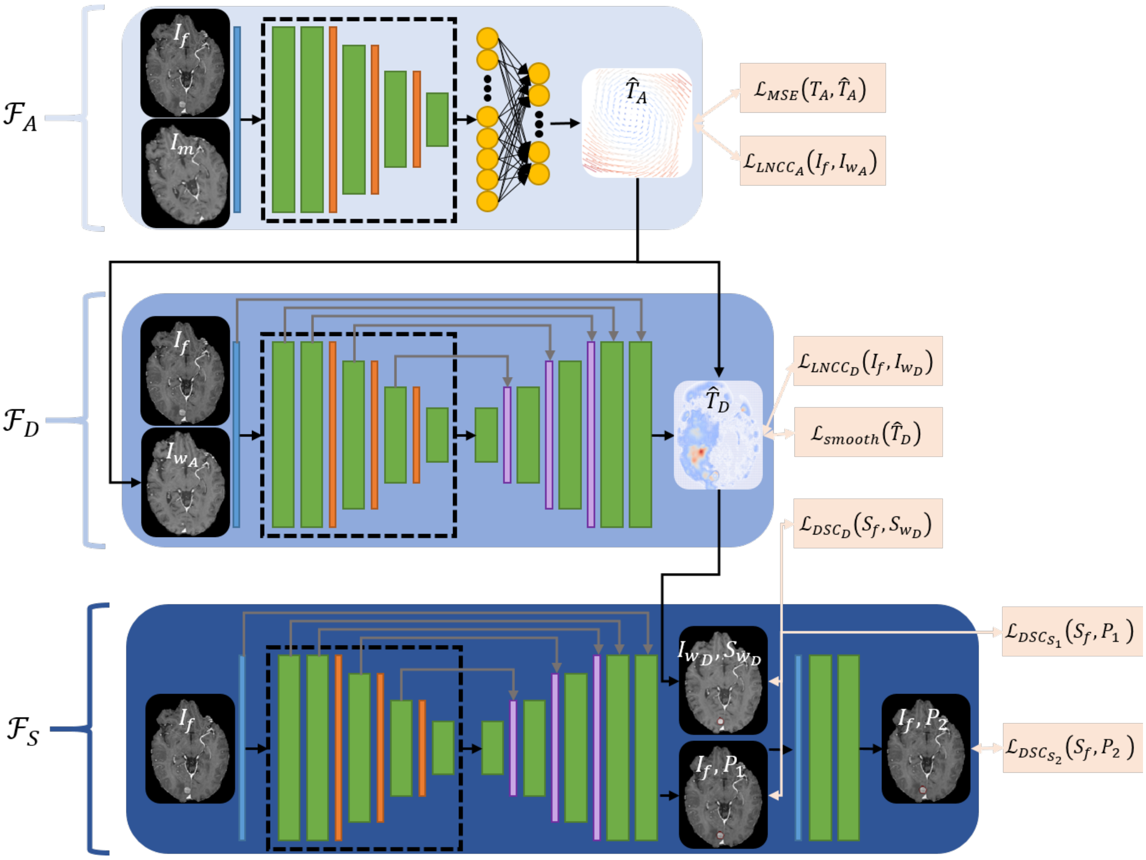

$\mathcal{F}_A$
I_f
I_m
$\hat{T}_A$
$\mathcal{L}_{MSE}(T_A, \hat{T}_A)$
$\mathcal{L}_{LNCC_A}(I_f, I_{w_A})$
$\mathcal{F}_D$
I_f
I_{w_A}
$\hat{T}_D$
$\mathcal{L}_{LNCC_D}(I_f, I_{w_D})$
$\mathcal{L}_{smooth}(\hat{T}_D)$
$\mathcal{L}_{DSC_D}(S_f, S_{w_D})$
$\mathcal{F}_S$
I_f
I_{w_D}, S_{w_D}
I_f, P_1
I_f, P_2
$\mathcal{L}_{DSC_{S_1}}(S_f, P_1)$
$\mathcal{L}_{DSC_{S_2}}(S_f, P_2)$

Figure 5-4: **Schematic showing the sequential structure of our proposed framework for joint image registration and segmentation.** First, the fixed image I_f, moving image I_m, and moving label S_m are passed to the affine registration network $\mathcal{F}_A$, which outputs the affinely warped image I_{w_A} and label S_{w_A}. Next, the fixed image I_f, affinely warped image I_{w_A}, and affinely warped label S_{w_A} are passed to the deformable registration network $\mathcal{F}_D$, which outputs the deformably warped image I_{w_D} and label S_{w_D}. Finally, the segmentation network $\mathcal{F}_S$ is used to segment I_f with the help of I_{w_D} and label S_{w_D} to output predicted segmentations P_1 (which does not use any prior time-point information) and P_2 (which uses prior time-point information). The encoder, which is shared between all three sub-networks, is shown in the dotted box. Light blue, blue, and dark blue are used to differentiate the three arms of the network.

5.4.2 Affine Registration Network $\mathcal{F}_A$ Architecture

Given input images I_f, I_m and S_m, the affine registration network $\mathcal{F}_A$ outputs $\hat{T}_A$, I_{w_A}, S_{w_A}= $\mathcal{F}_A(I_f, I_m, S_m)$, where $\hat{T}_A \in \mathbb{R}^{3\times4}$ represents the 3D affine transformation and $I_{w_A} = I_m \circ \hat{T}_A$ and $S_{w_A} = S_m \circ \hat{T}_A$ are the affinely warped moving image and label, respectively. $\mathcal{F}_A$ is composed of an opening convolution operation, the shared encoder, and a specialized affine transform decoding module. I_f and I_m are passed to the opening convolution, which uses a kernel size 7 with stride of 1. Increasing the kernel size from 3 to 7 for this opening convolution helps increase the effective receptive field (ERF) of the network, allowing for larger transformations to be learned. The decoding module is composed of two fully connected layers with dropout [224] of 0.15.

To train $\mathcal{F}_A$, we use a combination of two losses. First, we take the MSE loss between the true affine matrix T_A and the predicted affine matrix $\hat{T}_A$.

$$\mathcal{L}_{MSE}(T_A, \hat{T}_A) = \|T_A - \hat{T}_A\|_2^2 \tag{5.1}$$

Second, we utilize the unsigned local normalized cross correlation (LNCC) to measure the similarity between I_f and I_{w_A} [118, 124]. We define the local mean centered image $\overline{I}_f = I_f - \mu_{I_f}$, where μ_{I_f} is the convolved output of I_f and a kernel size 9 box filter. $\overline{I}_{w_A}$ is defined similarly. The LNCC is then given by:

$$LNCC(I_f, I_{w_A}) = \frac{\langle I_f - \mu_{I_f}, I_{w_A} - \mu_{I_{w_A}} \rangle^2}{\langle I_f - \mu_{I_f}, I_f - \mu_{I_f} \rangle \langle I_{w_A} - \mu_{I_{w_A}}, I_{w_A} - \mu_{I_{w_A}} \rangle} = \frac{\langle \overline{I}_f, \overline{I}_{w_A} \rangle^2}{\langle \overline{I}_f, \overline{I}_f \rangle \langle \overline{I}_{w_A}, \overline{I}_{w_A} \rangle} \tag{5.2}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product. LNCC ranges from 0 to 1, with 1 representing perfectly aligned images. To use as a loss function, we take the negative of the value.

$$\mathcal{L}_{LNCC_A}(I_f, I_{w_A}) = -LNCC(I_f, I_{w_A}) \tag{5.3}$$

5.4.3 Deformable Registration Network $\mathcal{F}_D$ Architecture

Given input images I_f, I_{w_A}, and S_{w_A}, the deformable registration network $\mathcal{F}_D$ outputs $\hat{T}_D$, I_{w_D}, $S_{w_D} = \mathcal{F}_D(I_f, I_{w_A}, S_{w_A})$, where $\hat{T}_D \in \mathbb{R}^{h \times w \times d \times 3}$ represents the deformable transformation and $I_{w_D} = I_{w_A} \circ \hat{T}_D$ and $S_{w_D} = S_{w_A} \circ \hat{T}_D$ are the (affinely and) deformably warped moving image and label, respectively. $\mathcal{F}_D$ is composed of an opening convolution operation, the shared encoder, and a specialized deformable transform decoding module. The decoding module is the inverse of the shared encoder, with trilinear upsampling layers in lieu of max pooling. Following standard U-Net approaches, we interleave skip connections from the encoder to the decoder.

To train $\mathcal{F}_D$, we use a combination of three losses. First, we measure the similarity between I_f and I_{w_D} as follows:

$$\mathcal{L}_{LNCC_D}(I_f, I_{w_D}) = -LNCC(I_f, I_{w_D}) \tag{5.4}$$

Next, to encourage the predicted DVF to be spatially smooth, we use the following second order bending energy penalty [115]:

$$\begin{aligned}\mathcal{L}_{smooth}(\hat{T}_D) = \sum \Bigg(& \left\| \frac{\partial^2 \hat{T}_D}{\partial x} \right\|_2^2 + \left\| \frac{\partial^2 \hat{T}_D}{\partial y} \right\|_2^2 + \left\| \frac{\partial^2 \hat{T}_D}{\partial z} \right\|_2^2 \\ & +2 \left\| \frac{\partial^2 \hat{T}_D}{\partial xy} \right\|_2^2 + 2 \left\| \frac{\partial^2 \hat{T}_D}{\partial xz} \right\|_2^2 + 2 \left\| \frac{\partial^2 \hat{T}_D}{\partial yz} \right\|_2^2 \Bigg)\end{aligned} \tag{5.5}$$

where spatial gradients are approximated via a second order finite difference. If we place too much weight on this penalty, the predicted DVF will be over-smoothed and will not adequately align I_f and I_{w_D}. Conversely, if we do not penalize enough, we may see physiologically unrealistic transformations such as folding or other discontinuities.

Finally, to add extra incentive to the network to learn how to accurately shrink or enlarge tumors (which will improve our downstream segmentation performance), we utilize the Dice Score Coefficient (DSC) [71]. Given two label maps p and q, the DSC measures how well they overlap as follows:

$$DSC(p, q) = \frac{2 \sum pq}{\sum p + \sum q} \tag{5.6}$$

DSC ranges from 0 to 1, with 1 representing perfect overlap. To use as a loss function, we take the negative of the value.

$$\mathcal{L}_{DSC_D}(S_f, S_{w_D}) = -DSC(S_f, S_{w_D}) \tag{5.7}$$

5.4.4 Segmentation Network $\mathcal{F}_S$ Architecture

Given input images I_f, I_{w_D}, and S_{w_D}, the segmentation network $\mathcal{F}_S$ outputs P_1, P_2 $= \mathcal{F}_S(I_f, I_{w_D}, S_{w_D})$, where P_1 and P_2 are pixelwise probability maps for likely brain metastases. $\mathcal{F}_S$ is composed of an opening convolution operation, the shared encoder, and a specialized segmentation module. This module works slightly differently from the prior two in that the input to the opening convolution is solely the fixed image. The output of the segmentation decoding module, which follows the same structure as the deformable decoding module, is P_1. As this part of the segmentation module is run solely using the fixed image, it does not incorporate any prior time-point information at this point. The second part of the segmentation module is a residual block [33] which fuses information from the current time-point (I_f and P_1) with information from the prior time-point (I_{w_D} and S_{w_D}) to output the final enhanced segmentation P_2.

To train $\mathcal{F}_S$, we apply DSC loss to both P_1 and P_2 as follows:

$$\mathcal{L}_{DSC_{S_1}}(S_f, P_1) = -DSC(S_f, P_1) \tag{5.8}$$

$$\mathcal{L}_{DSC_{S_2}}(S_f, P_2) = -DSC(S_f, P_2) \tag{5.9}$$

5.4.5 Pyramidal Architecture

In this section, we will briefly describe the pyramidal structure of our network architecture, a schematic of which is shown in figure 5-5. To begin, we use a L-level pyramid framework for both our affine and deformable registration networks, where we set $L = 3$ for this paper. For level $i \in \{1, 2, 3\}$ in the pyramid, the input images are downsampled by a factor 0.5^{L-i}. A forward pass through the pyramidal structure entails iteratively registering the images at from the coarsest scale (level $i = 1$) to the finest scale (level $i = 3$). More specifically, at pyramid level $i = 1$, we downsample images I_f and I_m by a factor of $0.5^{L-i} = 4$ to obtain coarse images I_{f_1} and I_{m_1}. These are passed through $\mathcal{F}_{A_1}$ to output a coarse affine transformation $\hat{T}_{A_1}$. At pyramid

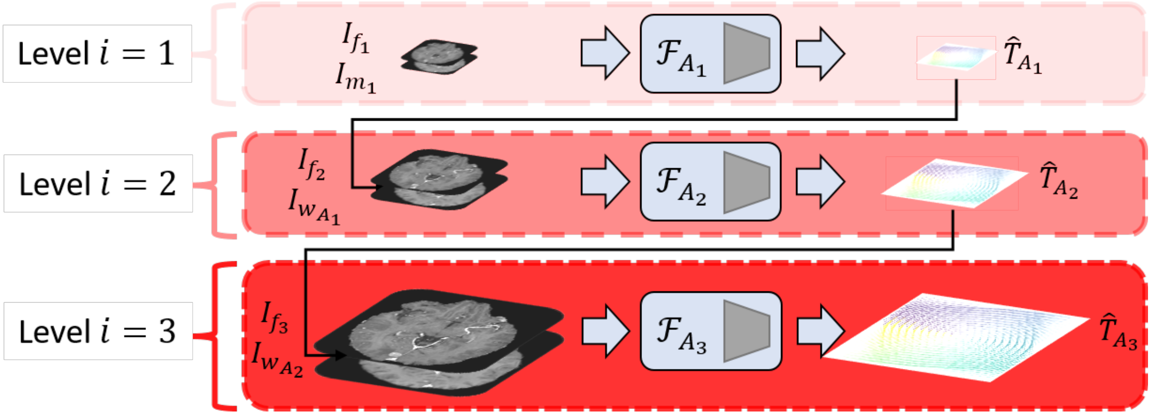


Figure 5-5: **Schematic showing the pyramidal structure of our proposed framework for affine image registration.** The predicted affine transformation $\hat{T}_A$ is iteratively refined by registering the images from the coarsest scale (level $i = 1$) to the finest scale (level $i = 3$). The pyramidal structure for deformable registration follows the same approach. Light red, red, and dark red are used to differentiate the three levels of the pyramid.

levels $i > 1$, we downsample images I_f and I_m by the appropriate scale factor to obtain images I_{f_i} and I_{m_i} and we warp I_{m_i} with the previously computed transform $\hat{T}_{A_{i-1}}$ to make $I_{w_{A_{i-1}}}$. I_f and $I_{w_{A_{i-1}}}$ are passed through $\mathcal{F}_{A_i}$ to output a refined affine transformation $\hat{T}_{A_i}$. The pyramidal structure for deformable registration follows the same approach.

As stated previously, the pyramidal sub-networks $\mathcal{F}_{A_i}$ and $\mathcal{F}_{D_i}$ for $i \in \{1, 2, 3\}$ share the same trainable parameters. The only difference between sub-networks is that level $i = 1$ uses two fewer and level $i = 2$ uses one fewer non-trainable max pooling and trilinear upsampling layer than does level $i = 3$, respectively. By carefully removing down-sampling and up-sampling layers for the coarser image resolutions, we ensure that all input and output dimensions match up.

5.4.6 Total Pyramidal Loss Function

To balance losses coming from different levels in the pyramid, we use the following formulation:

$$
\begin{aligned}
&\mathcal{L}_{total}(T_A, \hat{T}_A, I_f, I_{w_A}, I_{w_D}, T_D, S_f, S_{w_D}, P_1, P_2) = \\
&\sum_{i=1}^{L} 0.5^{L-i}(\mathcal{L}_{MSE}(T_A, \hat{T}_A) + \mathcal{L}_{LNCC_A}(I_f, I_{w_A}) + \mathcal{L}_{LNCC_D}(I_f, I_{w_D}) \\
&+ \mathcal{L}_{smooth}(T_D) + \mathcal{L}_{DSC_D}(S_f, S_{w_D}) + \mathcal{L}_{DSC_{S_1}}(S_f, P_1) + \mathcal{L}_{DSC_{S_2}}(S_f, P_2))
\end{aligned}
\tag{5.10}
$$

where γ controls how much to decrease the loss at coarser image scales, λ_1 controls the strength of $\mathcal{L}_{smooth}$, and λ_2 is a weighting hyperparameter to prevent $\mathcal{L}_{DSC_D}$ from overpowering $\mathcal{L}_{smooth}$ and resulting in spatially discontinuous deformations at tumor boundaries.

5.5 Cohort

We acquired a cohort of patients with clinically diagnosed BM from a retrospective database from Brigham and Women's Hospital (BWH). We selected adult patients with newly diagnosed BM who were undergoing stereotactic radiosurgery treatment from April 2004 to November 2014. This yields 148 patients with 885 time-points total. To train our registration and segmentation model, we divided this cohort on the patient level into training (100 patients; 617 time-points), validation (25 patients; 139 time-points), and testing (23 patients; 129 time-points) sets. To better understand how model performance varies as a function of lesion volume, each set was sub-divided into groups of consisting of micro ($< 25\text{mm}^3$), small ($\geq 25\text{mm}^3$ and $< 125\text{mm}^3$), medium ($\geq 125\text{mm}^3$ and $< 1000\text{mm}^3$), and large ($\geq 1000\text{mm}^3$) lesions. Table 5.1 details demographical statistics of these selected cohorts. All patient examinations were viewed and annotated in Slicer3D [197]. Segmentations were first manually segmented by a neuro-oncologist and then manually edited by a board-certified neuro-radiologist with 16 years' experience. These segmentations serve as the ground truth for our experiments.

Table 5.1: Patient demographic information for the selected brain metastases cohort.

Characteristic	Training Set	Validation Set	Test Set
Patient Characteristics			
No. of patients	100	25	23
No. of women*	68 (68)	13 (52)	13 (57)
Median age†	60 (52 - 66)	61 (54 - 71)	67 (60 - 73)
Examination Characteristics			
No. of MR examinations	617	139	129
No. of micro lesions	594	82	637
No. of small lesions	1245	138	465
No. of medium lesions	1070	141	156
No. of large lesions	522	128	72
Primary Cancer Type			
Lung*	51 (51)	10 (40)	16 (70)
Breast*	21 (21)	6 (24)	1 (4)
Melanoma*	17 (17)	7 (28)	3 (13)
Gastrointestinal*	3 (3)	1 (4)	2 (9)
Renal*	4 (4)	1 (4)	0 (0)
Other/Unknown*	4 (4)	0 (0)	1 (4)

* Data in parantheses are percentages.

† Data in parantheses are IQRs.

5.6 Results

Our goal is A) to accurately and efficiently compute affine and deformable registrations for brain tumor imaging data, and B) to improve segmentation of brain metastases by using prior time-point information. In this section, we quantitatively evaluate these goals.

5.6.1 Evaluation Approach/Study Design

We will evaluate registration performance as follows:

1. By computing LNCC between the fixed image I_f and the moved images I_{w_A} and I_{w_D}.

2. By computing DSC between the fixed label S_f and the moved labels S_{w_A} and

S_{w_D}.

Since ANTS and VXM are the most well-known and most rigorously validated classical and deep learning based registration methods, respectively, we will compare our method SPIRS to these two baselines. Namely, we will employ the Wilcoxon signed-rank test [225], the non-parametric analog to the paired t-test. Following recommendation from [124], we change the ANTS default parameters to use a step size of 0.25, Gaussian parameters of (9.0, 0.2), and three levels in the pyramid with at most 201 iterations each.

We will evaluate segmentation performance as follows:

1. By computing DSC, 95th percentile of Hausdorff distance (HD95) [226], and sensitivity of lesion detection between the fixed label I_f and the predicted segmentations P_1 and P_2.

For two arbitrary segmentations, Hausdorff distance is a distance metric which measures the furthest distance from any point on one boundary to its closest point on the other boundary. This is formulated as

$$HD(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B}\|a-b\|, \sup_{b\in B}\inf_{a\in A}\|a-b\|\right\} \tag{5.11}$$

where the 95th percentile of the HD is often used as it is a more robust metric unaffected by outliers.

We also assess sensitivity of lesion detection, which is given by:

$$Sensitivity = \frac{TP}{TP+FN} \tag{5.12}$$

where TP and FN are the number of true positives and false negatives, respectively. We hypothesize that using segmentation labels generated using prior time-point information (P_2) will significantly improve sensitivity of lesion detection when compared to using segmentation labels generated without any prior time-point information (P_1). To verify this, we will employ McNemar's test, a type of chi-squared test for paired data.

Statistical analyses were performed using Python 3.6.9 [169] and Matlab 2018b [227]. Statistically significant difference was set at $p \leq .05$.

5.6.2 Data Preprocessing

To mitigate variability between examinations stemming from potential differences in imaging protocols, we apply the following preprocessing steps. First, we resample all data to 1mm isotropic resolution and skull strip using antsBrainExtraction.sh (ANTs 2.3.1), using the OASIS atlas as the target template. To compensate for intensity inhomogeneity, we apply N4 bias correction [199] and normalize the image intensities to have zero mean, unit variance (based on non-zero intensity voxels only). Next, all volumes are tightly cropped to remove empty voxels outside the brain region. Finally, we affinely align all imaging time-points to their respective baseline scans via ANTS. In other words, all intra-patient imaging is affinely registered into the same space, allowing us to utilize a supervised training loss for the affine network. Individual images may have slightly varying sizes post skull-stripping and cropping. Indeed, while the median image size is $136 \times 169 \times 134$ voxels, the minimum size and maximum sizes are $121 \times 148 \times 116$ voxels and $160 \times 201 \times 159$ voxels, respectively. That is almost a factor of two more voxels in the maximum image size than the minimum. To account for these different image sizes and to ensure that our network architecture will fit into GPU memory, we resize all imaging to a uniform size of $128 \times 128 \times 128$ voxels. This resized image is approximately two thirds the volume of an image of median size. While this decrease may seem relatively large, qualitative evaluation indicates minimal loss in detail between the non-resized and resized images. Thus during training, the three levels of our pyramid correspond to images of size $32 \times 32 \times 32$ voxels, $64 \times 64 \times 64$ voxels, and $128 \times 128 \times 128$ voxels.

5.6.3 Implementation and Training Details

Our model SPIRS was implemented in DeepNeuro [228] with Tensorflow 2.10 backend [229]. The three hyperparameters γ, λ_1, and λ_2 in our total pyramidal loss function

$\mathcal{L}_{total}$ are set to 0.5, 0.5, and 0.1, respectively. These hyperparameter choices were determined experimentally. Full training details are described below.

Training is done using the SGD optimizer with decoupled weight decay [230] and we progressively decrease the learning rate via the following cosine decay schedule:

$$\eta_t = \eta_{min} + 0.5(\eta_{max} - \eta_{min})(1 + cos(\pi T_{curr}/T)) \tag{5.13}$$

where η_{max} is our initial learning rate (set to 1e-2), η_{min} is our final learning rate (set to 4e-5), T_{curr} is the current iteration counter, and T is the total number of iterations to train for (set to 250 epochs).

To mitigate overfitting, we apply weight decay of 4e-5 to all convolutional kernel parameters, leaving biases and scales un-regularized. The same cosine decay schedule is applied, where we set the final weight decay to be 2e-7. Furthermore, we apply real-time data augmentation during the training process. Specifically, we utilize random mirror axis flips about all three axes along with anisotropic scaling (0.9 to 1.1), rotations ($-20°$ to $20°$), shearing (-0.05 to 0.05), and translations (-20 to 20 pixels). Intensity augmentation in the form of gamma correction (.75 to 1.25) is used as well.

End-to-end training for our network is unstable due to the difficulty in balancing losses computed at different levels in the pyramid. To overcome this issue, we use a coarse-to-fine training approach. Since network parameters are shared between all levels in the pyramid, we instead begin by training only the coarsest level and successively add the other levels each time the model converges. In particular, we start by training the affine, deformable, and segmentation modules only at level $i = 1$. Since our 3D images are downsampled by a factor of 4 at this scale, we can fit a batch size of 32 into GPU memory, which allows us to train with batch normalization. After convergence, we fine-tune this model with levels $i = 1$ and $i = 2$ together. Since our images are now larger (only downsampled by a factor of 2 at this scale), we drop the batch size to 2 to ensure our model still fits into memory and we freeze batch statistics. This process is repeated one more time at the final level of the pyramid. A simple flowchart showing how SPIRS registers and segments a fixed and moving image is shown in figure 5-6. This training process took around 48 hours on a NVIDIA Tesla V100 32GB GPU.

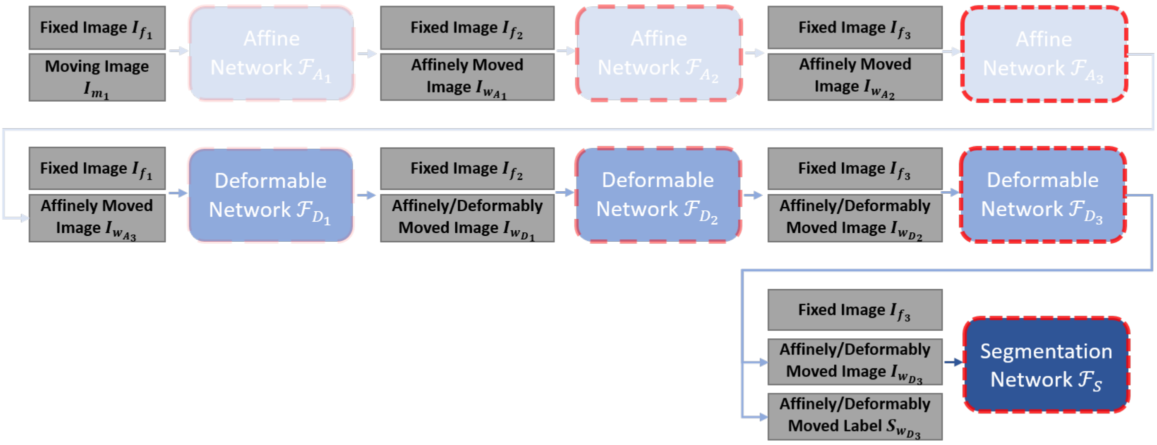

Fixed Image I_{f_1}
Moving Image I_{m_1}
Affine Network $\mathcal{F}_{A_1}$
Fixed Image I_{f_2}
Affinely Moved Image $I_{w_{A_1}}$
Affine Network $\mathcal{F}_{A_2}$
Fixed Image I_{f_3}
Affinely Moved Image $I_{w_{A_2}}$
Affine Network $\mathcal{F}_{A_3}$
Fixed Image I_{f_1}
Affinely Moved Image $I_{w_{A_3}}$
Deformable Network $\mathcal{F}_{D_1}$
Fixed Image I_{f_2}
Affinely/Deformably Moved Image $I_{w_{D_1}}$
Deformable Network $\mathcal{F}_{D_2}$
Fixed Image I_{f_3}
Affinely/Deformably Moved Image $I_{w_{D_2}}$
Deformable Network $\mathcal{F}_{D_3}$
Fixed Image I_{f_3}
Affinely/Deformably Moved Image $I_{w_{D_3}}$
Affinely/Deformably Moved Label $S_{w_{D_3}}$
Segmentation Network $\mathcal{F}_S$

Figure 5-6: **Flowchart showing the path through the SPIRS network.** Both the affine and deformable registration network is run at three levels. The segmentation network is only run at the finest level, since micro-metastases are down-sampled away at coarser image resolutions. Color coding follows from previous figures, with shades of blue representing different arms of the network and shades of red representing different levels of the pyramid.

5.6.4 Inference

Even though model training is done on resized images, inference is mainly done at the original resolution. There are two ways to perform inference at full resolution when a model is trained on smaller images.

First, a patch-based sliding-window approach can be adopted. In this method, patches the same size as the training patch (i.e. $128 \times 128 \times 128$ voxels) are passed to the network. These patches overlap the prior patch by a chosen amount, usually anywhere from 50% to 80%. We use a non-uniform averaging where more weight is given to the center of a patch than the extremities of a patch. At the end, all the patches are stitched and averaged together to produce a final full-size registration map. This process is extremely computationally expensive, since it requires many hundreds of forward passes through the network to get a full-size prediction out. The number of forward passes is a function of the patch overlap value. The higher the chosen patch overlap, the more time it will take to do a registration with the network. Lowering the patch overlap will expedite the process, but at the expense of registration accuracy.

The second way involves taking advantage of the convolutional nature of our network. The affine registration network contains two fully-connected layers at the end which output the 15 affine registration parameters. While convolutional layers do not require a fixed image size and can generate feature maps of any sizes, fully-connected layers on the other hand need to have fixed size/length input[231]. Due to the use of fully-connected layers in the affine registration network, we must run this component of the network with the same input resolution as during training (i.e. $128 \times 128 \times 128$ voxels). In order to utilize the predicted affine matrix $\hat{T}_A$ on full sized images, we must properly scale the affine transformation. Let h, w, and d refer to the height, width, and depth of a full sized image. The scaled affine transform is then given by:

$$Affine_{Full\_Size} = \begin{bmatrix} \frac{h}{128} & 0 & 0 \\ 0 & \frac{w}{128} & 0 \\ 0 & 0 & \frac{d}{128} \end{bmatrix} * \hat{T}_A * \begin{bmatrix} \frac{128}{h} & 0 & 0 \\ 0 & \frac{128}{w} & 0 \\ 0 & 0 & \frac{128}{d} \end{bmatrix} \tag{5.14}$$

where the first scale transform matrix is resizing the full size image to $128 \times 128 \times 128$ voxels and the second scale transform matrix is resizing back to the full size of $h \times w \times d$ voxels. This corrected affine matrix is used to warp the full size image, which can then be used in the deformable registration and segmentation networks, which are fully convolutional. As such, at test time we do not need to resize the images for these two sub-networks. Instead, we simply pass the full resolution images entirely at once to the networks. This process is highly efficient, taking less than 10 seconds per registration on GPU.

Test time augmentation (TTA) is a common practice for image classification and segmentation networks [232, 233, 234]. By applying certain transformations (such as flipping, translations, scaling, etc.) to the image at test time and then averaging network outputs across those different transformations, one can improve model performance by non-negligible amounts. As this practice has not been applied to image registration methods, we assess performance gain from using mirror axis flipped versions of the input. Specifically, we pass all eight mirror axis flipped versions of the fixed and moving images through the first level of our affine registration network. The resultant DVFs are un-flipped accordingly and averaged together. This averaged DVF is then used to warp the moving image for the second level of the affine pyramid. We note that during the training process, all warps are done via linear interpolation for simplicity. At test time, we replace linear interpolation with tri-cubic interpolation, which results in higher quality warped images. This process of collecting eight mirror axis flipped DVFs and averaging together is repeated a total of six times (three for the affine network and three for the deformable network). We note that this adds some additional computational burden as the full network is now run eight times.

5.6.5 Image Registration Results

Comparison Between Types of Inference

We begin by comparing inference using a patch-based sliding-window approach with a more simple approach wherein the full image is directly passed to the network. We report results for deformable registration in table 5.2.

Table 5.2: Comparing two different approaches to inference via SPIRS.

Method	LNCC[‡]	DSC[‡]	Mean $\|J\|$[‡]	$\|J\| \leq 0$[‡]
Patch	0.76 ± 0.05	0.71 ± 0.21	1.0021 ± 0.015	98.32 ± 296.24
Full	0.76 ± 0.06	0.70 ± 0.21	1.0 ± 0.015	185.06 ± 366.21

[‡] Data after ± is standard deviation.

We observe a slight, but not statistically significant, drop in mean LNCC and DSC when switching away from a patch-based approach. To quantify the smoothness of the deformation field, we compute the jacobian determinant of the predicted DVF. Negative values of the jacobian determinant indicate folding or other discontinuities that are physiologically impossible. We note an almost factor of two increase in this metric when switching from a patch-based approach ($p < 0.0001$). While this could be seen as worrisome, we note that an average of 185 voxels with negative jacobian determinant values accounts for less than 0.001% of all voxels in a 3D brain volume (which can have anywhere from two to five million voxels total). Due to the significant additional computation burden that the patch-based approach requires and the fact that LNCC and DSC are approximately unchanged, we utilize the full-scale inference technique for the rest of the analyses in this manuscript.

Test Time Augmentation Effects

We assess the performance gain of using TTA (i.e. averaging the outputs of eight mirror axis flipped versions of the input images). Specifically, we calculate the MSE between the true and predicted affine matrices, the LNCC between the fixed and moved images, and the MSE between fixed and moved images. These results are visualized in figure 5-7.

We observe that TTA provides better performance at almost every level of our pyramid. Specifically, the MSE between true and predicted affine matrix is lower when we use TTA at levels $i = 1$ ($p < 0.0001$), $i = 2$ ($p < 0.0001$), and $i = 3$ ($p < 0.01$). Similarly, the LNCC between the fixed and affinely moved images is higher when we use TTA at levels $i = 1$ ($p < 0.0001$) and $i = 2$ ($p < 0.01$), but not $i = 3$ ($p = 0.15$). This trend is reversed when comparing LNCC between the fixed and deformably moved images, where we do not find statistical significance for levels $i = 1$ ($p = 0.25$) or $i = 2$ ($p = 0.10$), but we do at $i = 3$ ($p < 0.01$). We note similar trends for when comparing MSE between fixed and moved images. Namely, we see that MSE is significantly lower for levels $i = 1$ ($p < 0.0001$) and $i = 2$ ($p < 0.05$) for the affine registration network, and for levels $i = 2$ ($p < 0.0001$) and $i = 3$ ($p < 0.01$) for the deformable registration network.

These results make intuitive sense. Affine registration is mainly accomplished at the coarser image scales, with very limited fine-tuning occurring at the finest image scale. Since the affine transform is unlikely to change drastically going from the second to the third level in the pyramid, we would expect to see little to no benefit to TTA at the third level. Conversely, deformable registration is used to align the high frequency

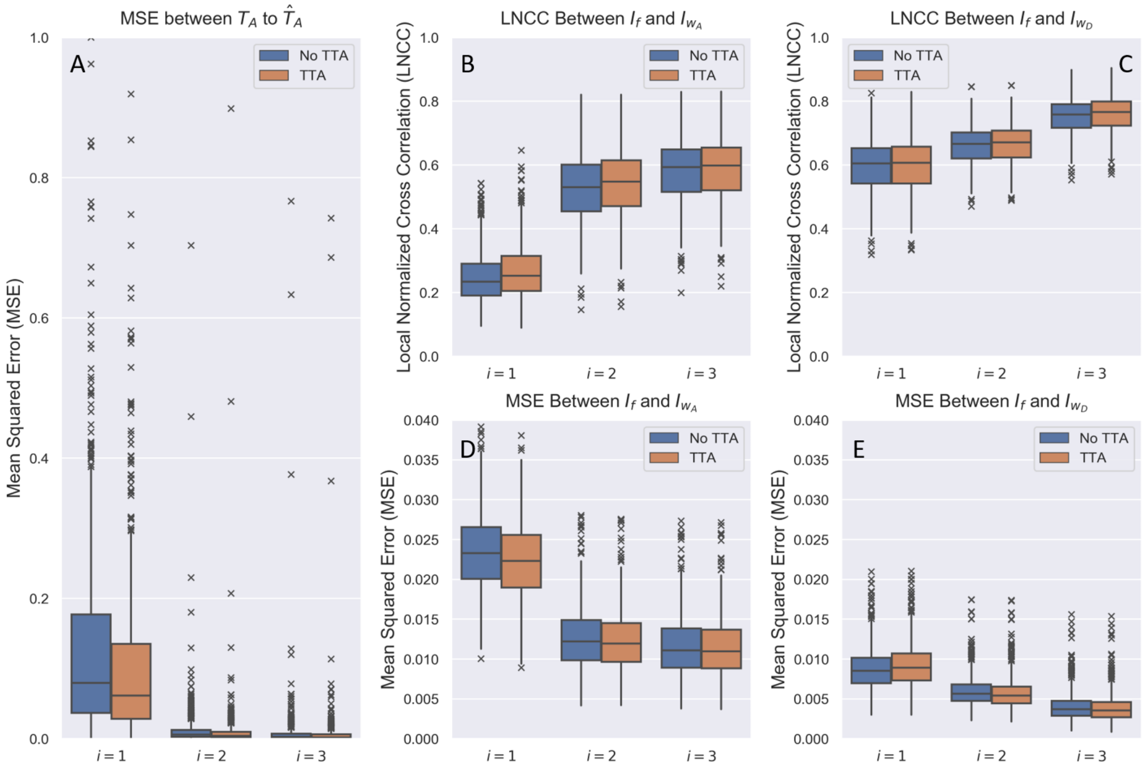


Figure 5-7: **SPIRS performance with and without test time augmentation.** (A) MSE between the true and predicted affine matrix. (B, C) LNCC between the fixed and affinely moved and deformably moved images, respectively. (D, E) MSE between the fixed and affinely moved and deformably moved images, respectively. We observe that test time augmentation (TTA) can provide minor, but noticeable improvements in both affine and deformable registration performance at each level in the pyramid.

components of an image. Thus the biggest deformations occur at the finest scale. Our results match this intuition, showing that TTA helps most at the finest level. This knowledge can be used to optimize computational burden by only applying TTA at select image scales where the benefit will be most appreciable.

For simplicity's sake and to ensure fair comparison to baseline methods, we will not be performing TTA for any of the analyses below.

Pyramidal Results

At each level $i \in \{1, 2, 3\}$ in the pyramid, we compute the LNCC between the fixed image I_{f_i} and the moved images $I_{w_{A_i}}$ and $I_{w_{D_i}}$. We find that LNCC markedly improves as we pass through the pyramid structures for both $\mathcal{F}_{A_i}$ and $\mathcal{F}_{D_i}$ (figure 5-8).

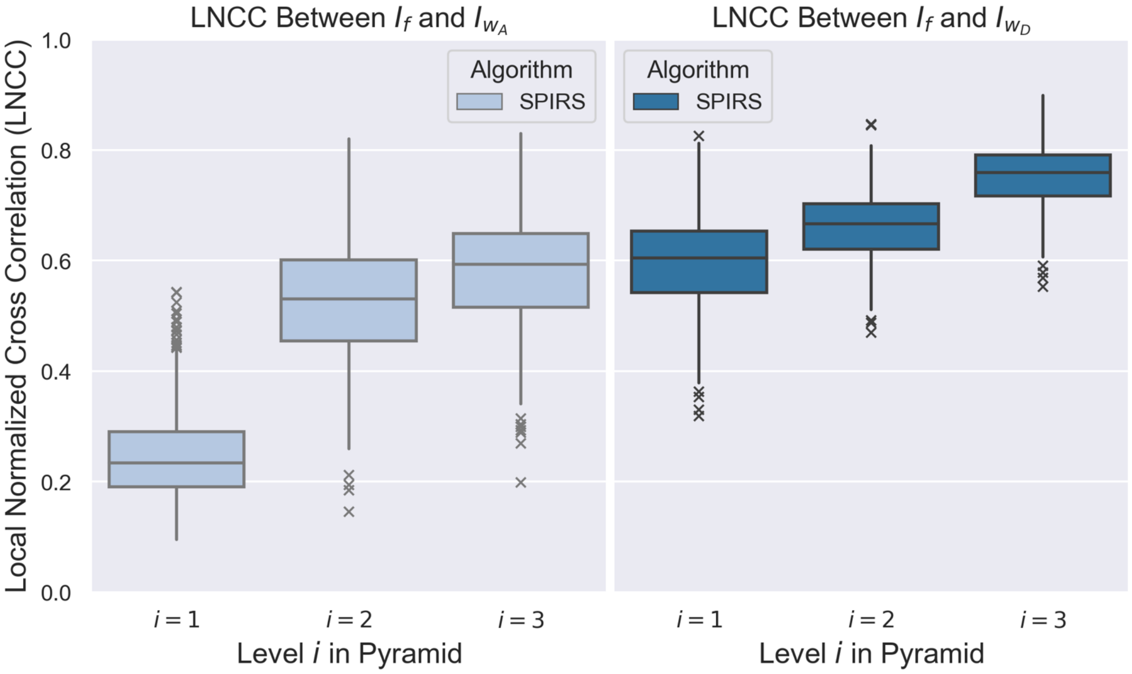


Figure 5-8: **Local normalized cross correlation at each level of the pyramid.** LNCC is computed for all intra-patient pairs of test set examinations at each level of the pyramid for affine and deformable registration. We observe that LNCC increases monotonically as we first go through the affine network and second go through the deformable network.

An example registration between two test set images is shown in figure 5-9 (A-H). Panel (A) shows the fixed image I_f and panel (B) shows the initially misaligned moving image I_m, which have an initial LNCC of 0.064. We note that the affine misalignment is purely 2D for the purpose of this figure, but emphasize that our network can register images fully in 3D. Panels (C-E) show the results of affine registration from the coarsest to the finest resolution along with the predicted transformations (which are visualized as a DVF). Due to the size of the figure, it is difficult to see the minute differences between $\hat{T}_{A_1}$, $\hat{T}_{A_2}$, and $\hat{T}_{A_3}$. Nevertheless, we note that the registration quality improved as evidenced by the LNCC, which rises to 0.110, 0.160, and 0.163 at levels 1, 2, and 3, respectively. Panels (F-H) show the results of deformable registration. Here, we can see striking differences between $\hat{T}_{D_1}$, $\hat{T}_{D_2}$, and $\hat{T}_{D_3}$, with LNCC rising to 0.398, 0.525, and 0.614, respectively. Qualitative analysis of this example registration confirms our quantitative analysis. We can see two major deformations that occur in the moving image. First, we notice that the tumor is significantly grown. Second, we notice that the right ventricle is shrunk to better match the appearance in the fixed image (and to better simulate the mass effect of the grown tumor). As this tumor located entirely in the right hemisphere of the brain, there is less deformation due to mass effect on the left side. This is visualized in our DVFs, which show much less deformation in the left hemisphere of the brain.

Comparison to Baselines

Next, we compare our method SPIRS against baseline registration methods ANTS and VXM (figure 5-10). Panels (A, C) compare LNCC and DSC between ANTS and SPIRS for affine registration; panels (B, D) compare LNCC and DSC between VXM, ANTS, and SPIRS for deformable registration. Since VXM only performs deformable registration (thus requiring images to be affinely aligned as a pre-processessing step), we affinely align all images via SPIRS before testing the three deformable registration algorithms to ensure fair comparison. Using LNCC and DSC as our metrics, we observe that SPIRS performs similarly to ANTS for affine registration ($p > 0.1$), and performs better than VXM and ANTS for deformable registration ($p < 0.0001$).

In figure 5-11, we show example deformable registrations between two pairs of test set images using VXM, ANTS, and SPIRS. SPIRS performs both qualitatively and quantitatively better than the other two methods. While all three methods can deformably register the large metastasis in the first row fairly well (though SPIRS subjectively does the best), we can see that only SPIRS can correctly deformably warp the smaller metastasis in the second row. In particular, VXM squeezes the metastasis into the midline, whereas ANTS does not warp it at all.

Ablation Study

Finally, to understand the effect of the pyramidal component of our registration network, we ran a small ablation study (table 5.3). When removing the pyramidal component of the network, we observe a decrease in median LNCC of 0.45 ($p < 0.0001$)

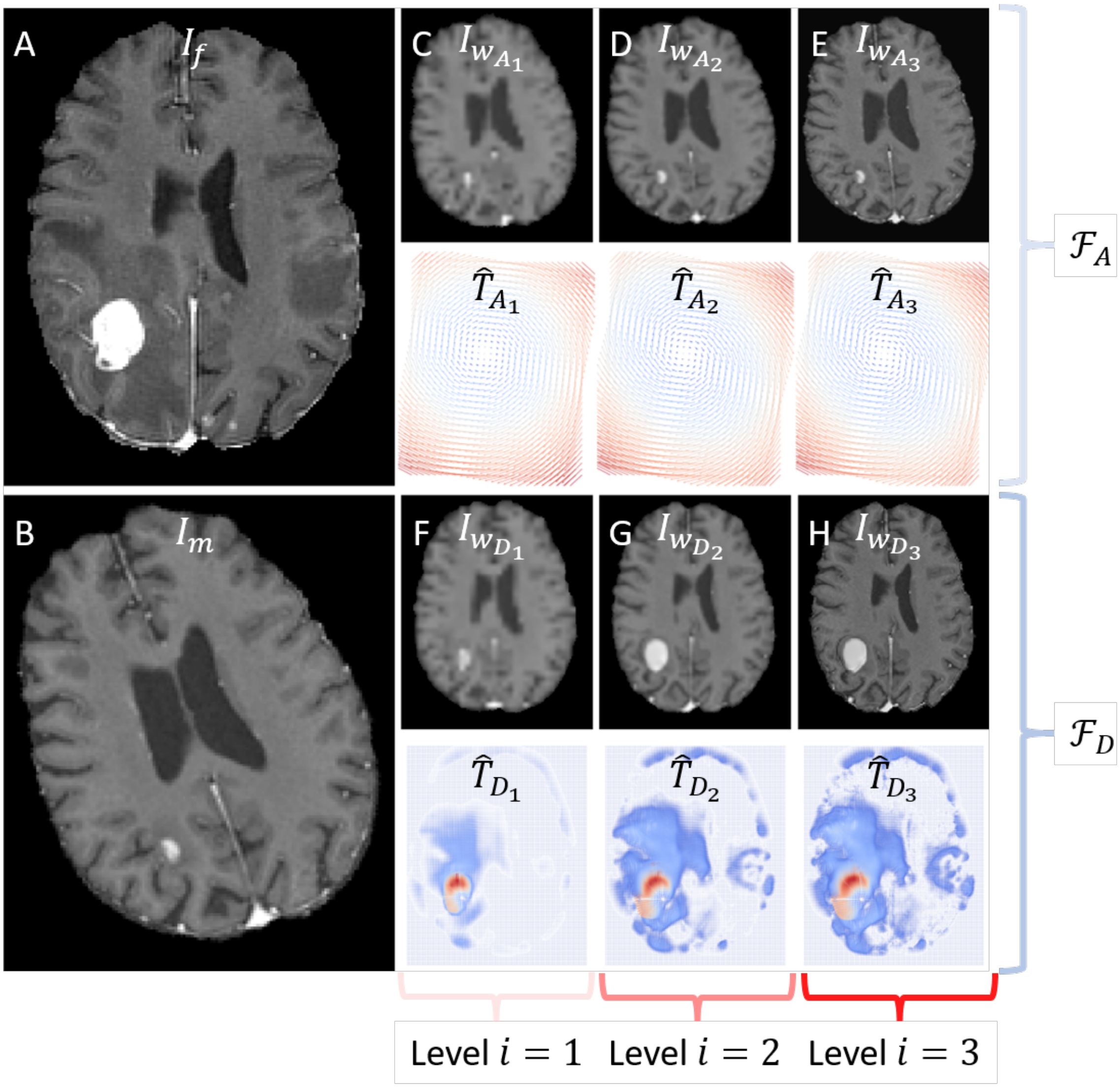


Figure 5-9: **An example affine and deformable registration via SPIRS.** A fixed image I_f (A) and a moving image I_m (B) from the test set are registered together via SPIRS. (C-E) show the results of pyramidal affine registration and (F-H) shows the results of pyramidal deformable registration.

and 0.04 ($p < 0.0001$) for affine and deformable registration, respectively. Similarly, we observe a decrease in median DSC of 0.25 ($p < 0.0001$) and 0.02 ($p < 0.0001$) for affine and deformable registration, respectively. This highlights the importance of using a multi-scale approach in order to guarantee optimal results, especially in the case of affine registration.

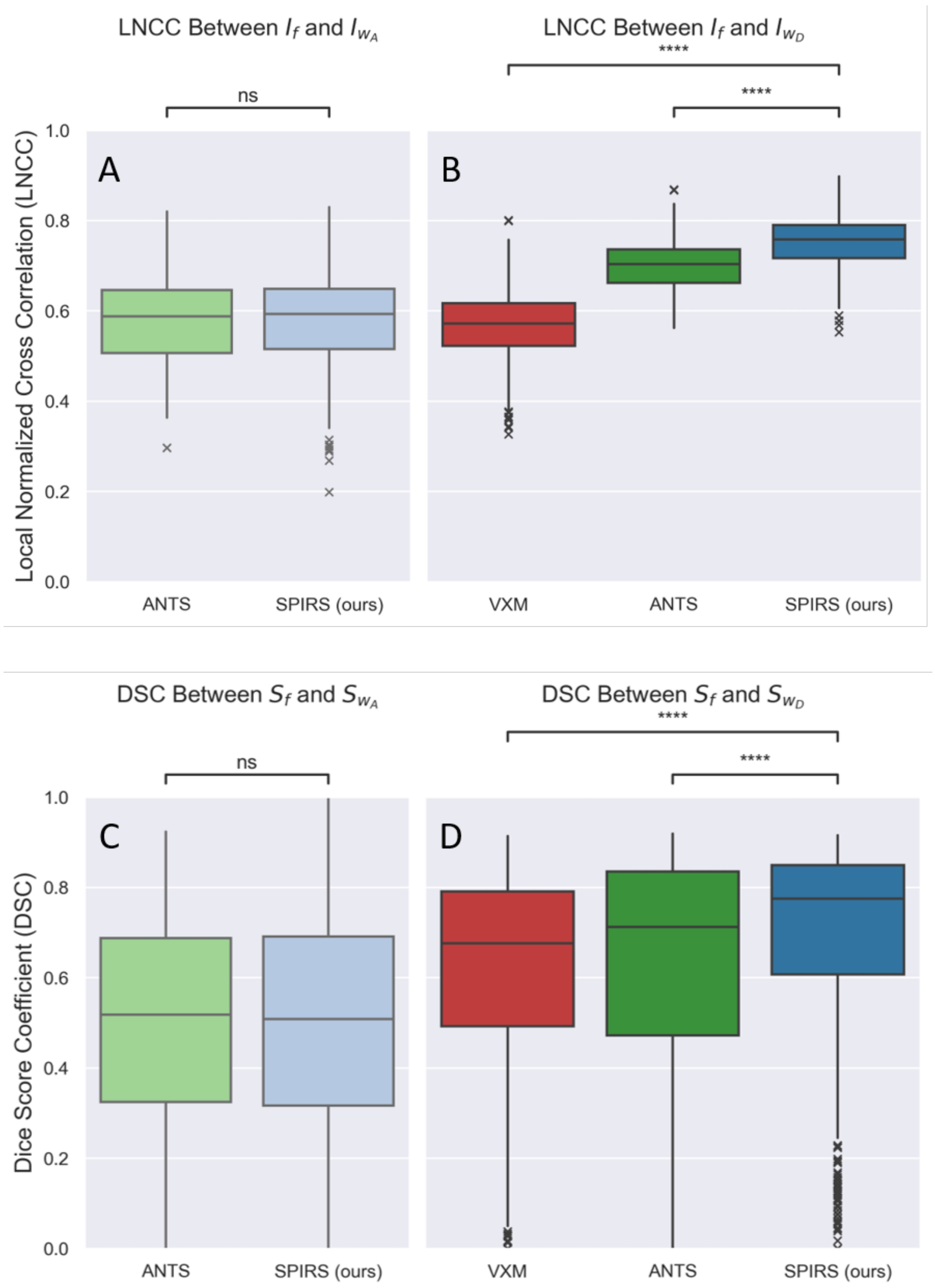


Figure 5-10: **Quantitative comparison between our method SPIRS and baseline methods VXM and ANTS.** (A, B) LNCC for affine and deformable registration. (C, D) DSC for affine and deformable registration.

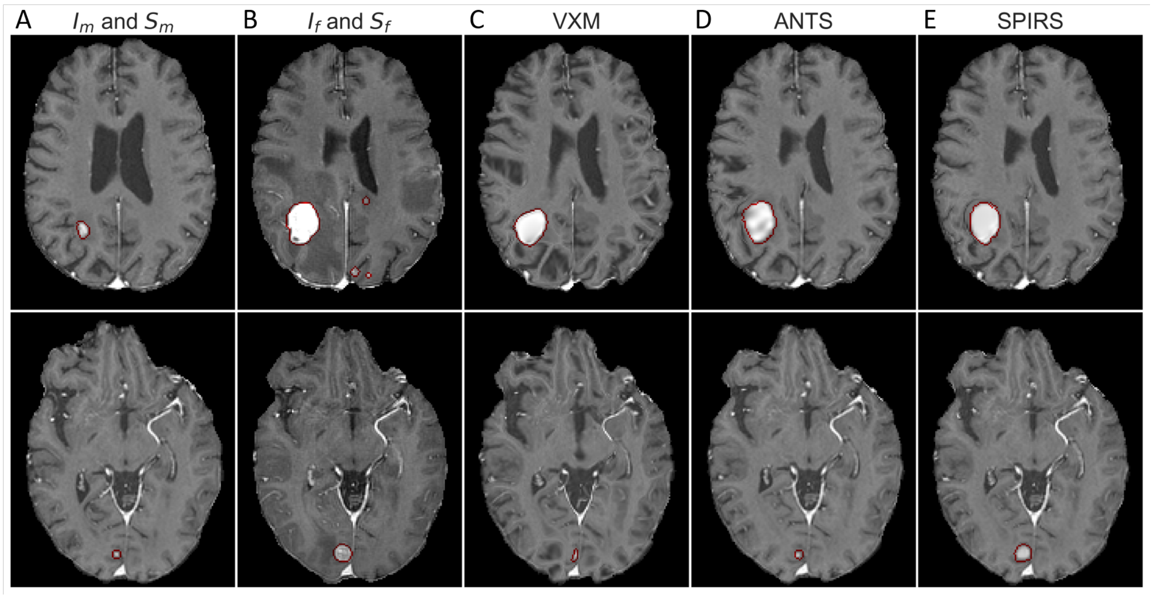


Figure 5-11: **Comparison of deformable registration quality between SPIRS, VXM, and ANTS.** (A) Moving image and moving label. (B) Fixed image and fixed label. (C-E) show deformable registration via VXM, ANTS, and SPIRS, respectively. We observe that SPIRS qualitatively and quantitatively performs the best, especially for smaller metastases.

Table 5.3: Ablation study to assess effect of pyramidal registration scheme on median LNCC and DSC.

Pyramidal	Transformation Type	LNCC[†]	DSC[†]
	Affine	0.14 (0.11 – 0.17)	0.26 (0.11 – 0.44)
✓	Affine	0.59 (0.52 – 0.65)	0.51 (0.32 – 0.69)
	Deformable	0.68 (0.64 – 0.71)	0.76 (0.72 – 0.79)
✓	Deformable	0.72 (0.53 – 0.82)	0.78 (0.61 – 0.85)

[†] Data in parantheses are IQRs.

5.6.6 Image Segmentation Results

We report results for the segmentation of BM split on the examination and lesion level with and without using prior time-point information in table 5.4 and is shown in figure 5-12. We observe minor improvements in DSC and HD95 at the examination level (but these are not statistically significant). Indeed, looking at figure 5-12(C), the two boxplots seem almost identical. When checking the difference in DSC between a predicted segmentation that uses prior information and one that does not in panel (A), we note two things. First, we observe that using prior information almost never reduces DSC. In other words, our model does not harm performance. Second, we observe that that any large increase in examination level DSC is concentrated in a small fraction of cases. Such results may imply that our method does not work as expected, but examination level data does not tell the full story. When looking at lesion level statistics, we observe a large increase in DSC (and accordingly large decrease in HD95), which are both statistically significant ($p < 0.0001$). Lesion level sensitivity also rises by 11% ($p < 0.0001$).

Table 5.4: Examination and lesion level median DSC, HD95, and sensitivity with and without using prior time-point information.

Level	Prior Info	DSC†	HD95†	Sensitivity*
Examination Level		0.84 (0.75 - 0.90)	1.41 (1.0 – 2.73)	N/A
Examination Level	✓	0.85 (0.75 - 0.90)	1.21 (1.0 – 1.93)	N/A
Lesion Level		0.17 (0.0 – 0.67)	3.64 (1.17 – inf)	53 (50 - 55)
Lesion Level	✓	0.4 (0.0 – 0.69)	1.97 (1.0 – inf)	64 (61 - 66)

* Data in parantheses are percentages.
† Data in parantheses are IQRs.

To identify where the improvement is coming from, we report results further subdivided by lesion size in table 5.5. We observe significant improvement in sensitivity of lesion detection for micro and small sized lesions when incorporating prior time-point information ($p < 0.0001$). Specifically, we detect 15% more micro-lesions (amounting to 92 fewer missed detections), and we detect 10% more small-lesions (amounting to 45 fewer missed detections). The sensitivity for medium and large sized lesions remains unchanged. Since we only detect 27% of micro-metastases without using prior information and 42% with using prior information, median DSC and HD95 are 0.0 and inf, respectively. Nonetheless, the 15% increase in sensitivity leads to significant improvement in these two metrics ($p < 0.0001$). We note a similar trend in DSC and HD95 for small metastases ($p < 0.0001$).

Examples of newly detected micro-metastatic lesions (all with ground truth volume

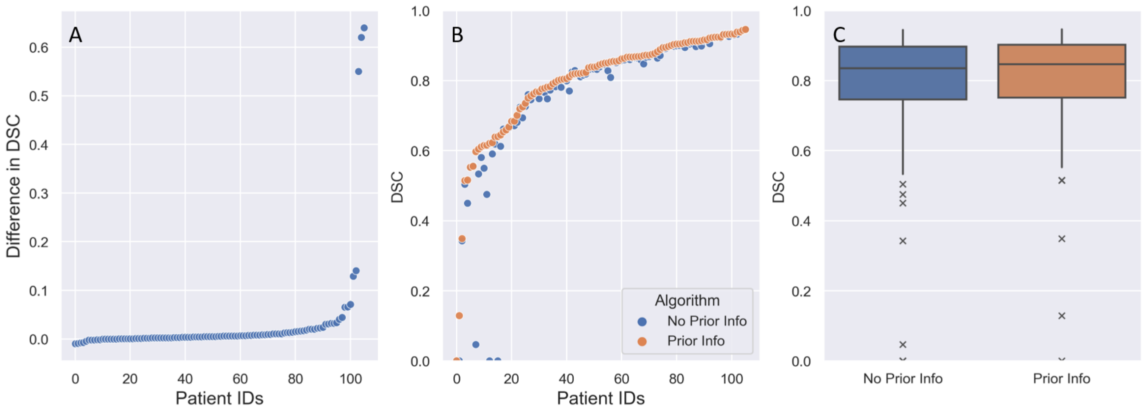


Figure 5-12: **Difference in segmentation performance of SPIRS with and without using prior time-point information.** (A) Difference in DSC with and without using prior time-point information. This shows that using prior information does not harm segmentation performance. Rather, for a small fraction of cases it can be highly beneficial. (B) Shows the same information but in the form of a paired scatter plot. (C) Shows the same information but in the form of a paired box plot.

Table 5.5: Median DSC, HD95, and sensitivity with and without using prior time-point information split by lesion size.

Lesion Size	Prior Info	DSC[†]	HD95[†]	Sensitivity[*]
micro		0.0 (0.0 - 0.21)	inf (2.12 – inf)	27 (24 - 31)
micro	✓	0.0 (0.0 - 0.44)	inf (1.41 – inf)	42 (38 - 46)
small		0.53 (0.0 – 0.68)	1.78 (1.0 – inf)	70 (66 - 74)
small	✓	0.57 (0.16 – 0.69)	1.41 (1.0 – 3.98)	80 (76 - 84)
medium		0.78 (0.74 – 0.84)	1.0 (1.0 – 1.41)	97 (92 - 99)
medium	✓	0.79 (0.74 – 0.85)	1.0 (1.0 – 1.41)	97 (92 - 99)
large		0.90 (0.88 – 0.92)	1.0 (1.0 – 1.41)	100 (93 - 100)
large	✓	0.91 (0.89 – 0.92)	1.0 (1.0 – 1.41)	100 (93 - 100)

[*] Data in parantheses are percentages.
[†] Data in parantheses are IQRs.

$< 10\text{mm}^3$) are shown in figure 5-13. Panel (A) shows an example of a patient with only one metastasis. In this case, this lesion is missed completely when not using prior time-point information, leading to an examination (and lesion) level DSC of 0.0. Even when using prior time-point information, we note that DSC only reaches slightly above 0.60. The segmentation is subjectively perfect, but minor differences between the ground truth and predicted labels can lead to big changes in DSC for small regions. In this case, a few voxel difference between the ground truth and predicted labels leads to a DSC of 0.60, even though the predicted segmentation label is subjectively better than the ground truth. Panels (B) and (C) show examples of cases with multiple metastases of different sizes. In these cases, due to the fact that there exists at least one more significantly larger metastasis which can be detected without prior information, the difference in examination level DSC is close to negligible ($< 2\%$). Cases like this highlight the need to look at lesion level statistics when there exist variability in metastases sizes.

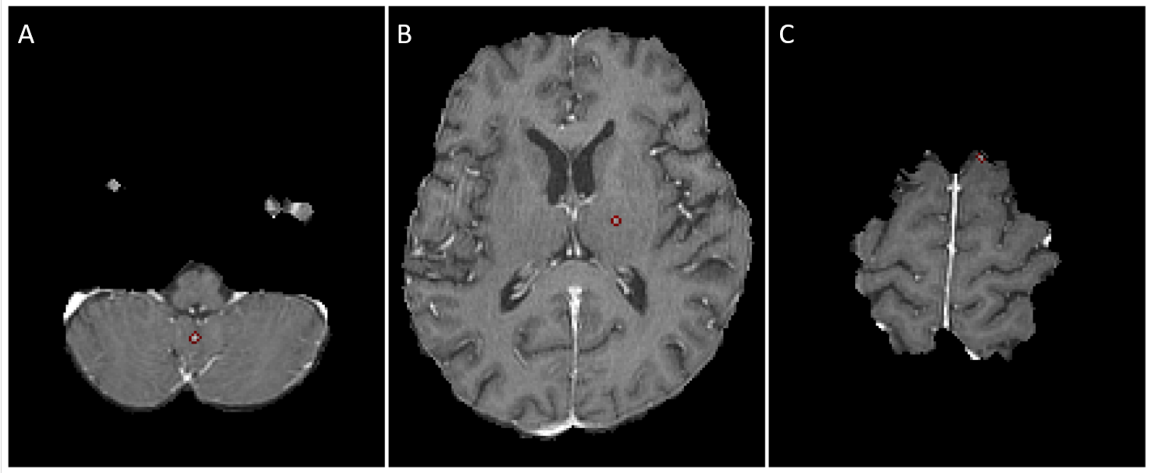


Figure 5-13: **Examples of SPIRS segmentation of brain micro-metastases.** Three different test set patients with automatically segmented micro-metastases (outlined in red) that are missed if not using prior time-point information.

5.7 Discussion

Automated segmentation of BM is challenging machine learning task, with current segmentation algorithms exhibiting poor performance for micro-metastatic lesions. Patients undergoing active treatment will require regular follow-up imaging scans for the purpose of treatment response assessment, and current segmentation approaches are likely to make the same mistakes repeatedly (e.g. micro-metastatic lesion missed at baseline is missed again at time-point 1). Instead of segmenting each image independently of each other, we propose to utilize the prior time-point imaging as a means to improve segmentation of the new time-point. To that end, we developed SPIRS, a method to affinely and deformably register a prior time-point image (with known ground truth segmentation) to a new time-point image for the purpose of improving segmentation performance on this new image. While we focus on BM for this paper, we emphasize that our model architecture is generalizable to other challenging medical segmentation problems where longitudinal imaging data is present.

In our experimental studies, we compare registration via SPIRS to registration via ANTS and VXM and the comparative analysis indicates that SPIRS performs equivalently for affine registration and performs better for deformable registration. We attribute this improvement in performance to the fact that most algorithms are developed for normal anatomy and are not well equipped to model large radial deformations like we see for tumor growth/shrinkage. Indeed, VXM performed the worst since none of the data it was trained on had any tumors. Using our method, we subsequently show that we can drastically decrease the number of missed detections of BM when we utilize prior time-point information. This has numerous clinical implications.

First, we can reduce the annotation burden on clinicians by decreasing the number of mistakes that must be corrected each time a patient comes in for follow-up imaging. This will help streamline clinical workflows and enable the clinician to spend more time working on important downstream tasks such as treatment response assessment. Next, our approach can help make longitudinal patient analysis more consistent. Due to a multitude of factors, patient follow-up imaging will occasionally be read and interpreted by a different radiologist. Not only can this can lead to inter-rater variability, where a lesion may be identified on one visit but not the other, but it may also have significant effects on the categorization of treatment response (e.g. accidentally assigning partial response (PR) instead of stable disease (SD)). Our method can help prevent such issues by ensuring that lesions that were identified in prior time-points are correctly carried forwards to new time-points. Third, many BM specific clinical trials are run independently at differing institutions ([235, 236, 237]).

In order to accurately assess treatment efficacy across multiple institutions and clinical trials, a standardized non-volumetric measurement system known as the response assessment in neuro-oncology (RANO) criteria is used ([179]). The RANO criteria has been shown to lead to higher amounts of inter- and intra-rater variability and has significantly poorer repeatability and consistency compared to true volumetric measurements ([187, 204]). Showing promising results for segmentation of BM, our approach is a step towards replacing RANO with volumetric tumor burden.

There are also certain interesting future directions in tumor growth modeling. Biologically inspired models of cell diffusion can be used to mathematically predict the dynamics of tumor growth [238, 239, 240]. As mentioned, our model not only warped the tumors properly, but was able to accurately warp the ventricles and other nearby structures to account for the tumor growth. These predicted DVFs could be used to model future tumor growth for a patient directly from imaging data without the need for an underlying model of tumor biology.

5.8 Limitations

There are a few limitations of our work. First, our dataset was collected retrospectively from a single institution. Model performance when used in a prospective manner is currently uncertain. Moreover, variations in scanner settings and MRI parameters between institutions can affect performance, and future work will entail validating our approach on a larger multi-site dataset. Second, our approach relies on the existence of high quality prior time-point segmentations. If the radiologist that interpreted the prior time-point missed a lesion or segmented a false positive, these mistakes will most likely be inadvertently carried through to the new time-point. Third, we did not run any exhaustive grid searches for the hyper-parameters in our approach. Minor improvements to both the registration and segmentation may be achieved through sophisticated hyper-parameter tuning. We also did not assess the effect of using higher resolution imaging during the training process, instead using resized images of size $128 \times 128 \times 128$ voxels. Further improvements may be realized by working at full resolution. Finally, we note that our algorithm could be susceptible to severe imaging artifacts. A human radiologist can easily account for image artifacts such as those stemming from patient motion and choose to throw away unreasonable data. On the other hand, our algorithm may fail catastrophically without any warning.

5.9 Future Directions

Deformable registration methods like VXM require that images be affinely aligned as a pre-processing step. In this manuscript, we developed a method for joint affine and deformable image registration, removing the need to initially co-register images. To the best of our knowledge, this is the first modern attempt to tackle this problem, with prior approaches being out-of-date and sub-optimal in different ways.

Looking to the future, we will apply our method to normal anatomy imaging, rigorously and fairly comparing to a wide range of classical [241, 119] and baseline DL based registration methods [124, 125, 242, 243]. To ensure optimal performance of our new model, we will make a few changes to our current model.

First, we will change the affine registration network to directly predict an affine matrix as opposed to predicting a set of affine parameters and converting those into an affine matrix. As matrix multiplication is not commutative, switching the order of transforms will affect the end result. Our current model is trained to sequentially apply rotation, scale, shear, and translation, in that order. As such, it is not able to model a transformation that might perform shearing before rotation. By directly regressing the affine matrix, this will enable us to learn a less constrained set of affine transforms. We will also increase the data augmentation we use so that our model learns more intense affine transforms. Specifically, we plan to utilize random mirror axis flips about all three axes along with anisotropic scaling (0.75 to 1.25), rotations (-45° to 45°), shearing (-0.20 to 0.20), and translations (-30 to 30 pixels). Heavier intensity augmentation in the form of gamma correction (.65 to 1.35) and random bias field addition will be used as well. For our previous model, we did not apply random elastic augmentations during training, but we will experiment with these types of augmentations in the future. Additionally, we will modify the affine registration architecture to remove the flattening layer that occurs right before the fully-connected layers. The flatten layer (along with the fully-connected layers) lock the image input size of the model, which is sub-optimal for our purposes. We will instead re-parameterize the fully-connected layers as convolutional layers. Note that a fully-connected layer is the same as a convolution with kernel size equal to the image size and valid padding (i.e. the kernel does not move). By applying this re-parameterization, our network can accept any input image size as input, allowing our affine registration network to run at full-scale resolution.

Besides making modifications to the affine network, there are also a host of other modifications to the training procedure. First, we will modify the LNCC loss function to use different window sizes at different levels in the pyramid. This will ensure that the local cross-correlation is dependent on window sizes that are proportionally equivalent to the image size at that level. Second, while the convolutional layers in the encoder will remain shared across networks and levels, we will make batch normalization distinct between the affine and deformable registration networks. Batch normalization is extremely dataset and task dependent, and we expect better performance when train batch statistics separately on the affine and deformable networks. Third, we will update the smoothing loss on the deformation field. As noted previously, our network produces a small, but non-zero, number of areas with negative jacobian determinant values. A common idea in machine learning is to optimize for the metric you want to maximize (or minimize). In this case, adopting a first derivative or second derivative smoothing penalty on the deformation field is a proxy for reducing negative jacobian determinant values. Instead of using these proxy methods, we will instead directly penalize the jacobian determinant, which we expect will remove all discontinuities in the predicted deformation field. Fourth, we will update our training procedure to include a consistency loss. Specifically, we will train the model to predict both the forwards and backwards transforms, and enforce that the composition of the forwards and backwards transforms are identity. This has the added benefit of allowing us to output both transforms simultaneously. Finally, since we are going to be training on normal anatomy imaging, we can use anatomical labels to improve registration quality. VXM has shown up to a 3% dice score improvement when using label masks, and we expect similar performance benefits.

5.10 Conclusion

In conclusion, we propose a novel DL based framework for joint image registration and segmentation of brain metastases on MRI. Our approach can be used to affinely and deformably co-register two intra-patient examinations and segment the metastatic lesions on the newer time-point. This tool may be used to reduce annotation burden for the clinician and improve sensitivity of detection of micro-metastatic brain lesions. Future work includes further improving the registration method and the initial segmentation network. We may also consider applying this approach to other relevant medical applications.

Chapter 6

Conclusion

Image segmentation plays a vital role in the clinical workflow, but a plethora of challenges prevents its widespread adoption and use. The research presented in this thesis presents methods to automate image segmentation and shows how having highly accurate automated segmentations can improve downstream clinical tasks.

In chapter 3, I developed the first fully automated deep learning based system for adrenal gland segmentation and classification. With the high inter-reader variability present for manual detection of adrenal masses, there is risk of accidentally missing clinically relevant disease. I envision that this model will inform radiologists, providing pre-reads of all imaging and thus, enabling more reproducible and repeatable detection.

In chapter 4, I developed a pipeline for longitudinal response assessment for brain metastases patients. Since the RANO-BM criteria does not quantify lesions below 10 mm, it is providing potentially incomplete depictions of tumor progression. I envision my pipeline for automatic volumetric quantification to enable radiologists to focus on more complex and important patient management tasks, rather than segmenting images all day long.

In chapter 5, I developed a novel joint image registration and segmentation framework to improve the segmentation of micro-metastases. By using prior time-point information, we can reduce future annotation burden for the clinician. In the prior two chapters, neural network based models were informing the radiologist. In this scenario, the causation is reversed, with the radiologist informing the neural network.

And that leads me to the final takeaway point of this dissertation. There has to be synergy between healthcare professions and deep learning based models. Specifically, there has to be an understanding of when models fails and why they fail. And if there is that understanding, then a positive feedback cycle is generated, where radiologists can improve the models, and the models can improve radiology.

Bibliography

[1] Bercovich, E. & Javitt, M. C. Medical imaging: From roentgen to the digital revolution, and beyond. *Rambam Maimonides Med J* **9** (2018).

[2] World health organization (2016). URL `https://www.who.int/news-room/feature-stories/detail/to-x-ray-or-not-to-x-ray-`. Accessed on April 13, 2023.

[3] Rogers, W. *et al.* Radiomics: from qualitative to quantitative imaging. *The British Journal of Radiology* **93**, 20190948 (2020). URL `https://doi.org/10.1259/bjr.20190948`. PMID: 32101448, `https://doi.org/10.1259/bjr.20190948`.

[4] Neuman, M. I. *et al.* Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children. *J Hosp Med* **7**, 294–298 (2011).

[5] Sakurada, S. *et al.* Inter-rater agreement in the assessment of abnormal chest x-ray findings for tuberculosis between two asian countries. *BMC Infectious Diseases* **12**, 31 (2012).

[6] Rajaraman, S., Sornapudi, S., Alderson, P. O., Folio, L. R. & Antani, S. K. Analyzing inter-reader variability affecting deep ensemble learning for covid-19 detection in chest radiographs. *PLOS ONE* **15**, 1–32 (2020). URL `https://doi.org/10.1371/journal.pone.0242301`.

[7] Hsieh, S. S. *et al.* Understanding reader variability: A 25-radiologist study on liver metastasis detection at ct. *Radiology* **306**, e220266 (2023). URL `https://doi.org/10.1148/radiol.220266`. PMID: 36194112, `https://doi.org/10.1148/radiol.220266`.

[8] Pitman, A. G. *et al.* Intrareader variability in mammographic diagnostic and perceptual performance amongst experienced radiologists in Australia. *Journal of medical imaging and radiation oncology* **55**, 245–251 (2011).

[9] Hietikko, R. *et al.* Expected impact of MRI-related interreader variability on ProScreen prostate cancer screening trial: a pre-trial validation study. *Cancer imaging : the official publication of the International Cancer Imaging Society* **20**, 72 (2020).

[10] Hesamian, M. H., Jia, W., He, X. & Kennedy, P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of digital imaging* **32**, 582–596 (2019).

[11] Yepes-Calderon, F. & Gordon McComb, J. Manual segmentation errors in medical imaging. proposing a reliable gold standard. In Florez, H., Leon, M., Diaz-Nafria, J. M. & Belli, S. (eds.) *Applied Informatics*, 230–241 (Springer International Publishing, Cham, 2019).

[12] Sharma, N. & Aggarwal, L. M. Automated medical image segmentation techniques. *Journal of medical physics* **35**, 3–14 (2010).

[13] Pednekar, G. V. *et al.* Image Quality and Segmentation. *Proceedings of SPIE–the International Society for Optical Engineering* **10576** (2018).

[14] Zhang, L. *et al.* Disentangling human error from the ground truth in segmentation of medical images (2020). 2007.15963.

[15] Liu, X., Song, L., Liu, S. & Zhang, Y. A review of deep-learning-based medical image segmentation methods. *Sustainability* **13** (2021). URL https://www.mdpi.com/2071-1050/13/3/1224.

[16] Zhang, D. *et al.* Deep learning for medical image segmentation: Tricks, challenges and future directions (2022). 2209.10307.

[17] Avanzo, M. *et al.* Machine and deep learning methods for radiomics. *Medical Physics* **47**, e185–e202 (2020). URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13678. https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13678.

[18] Beers, A. *et al.* Sequential neural networks for biologically informed glioma segmentation. In Angelini, E. D. & Landman, B. A. (eds.) *Medical Imaging 2018: Image Processing*, vol. 10574, 1057433. International Society for Optics and Photonics (SPIE, 2018). URL https://doi.org/10.1117/12.2293941.

[19] Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (The MIT Press, 2016).

[20] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

[21] Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks* **4**, 251–257 (1991). URL https://www.sciencedirect.com/science/article/pii/089360809190009T.

[22] Kidger, P. & Lyons, T. Universal approximation with deep narrow networks (2020). 1905.08539.

[23] Singh, A., Thakur, N. & Sharma, A. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1310–1315 (2016).

[24] Ouali, Y., Hudelot, C. & Tami, M. An overview of deep semi-supervised learning (2020). 2006.05278.

[25] Cao, K., Brbic, M. & Leskovec, J. Open-world semi-supervised learning (2022). 2102.03526.

[26] Chen, H. *et al.* An embarrassingly simple baseline for imbalanced semi-supervised learning (2022). 2211.11086.

[27] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations (2020). 2002.05709.

[28] He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning (2020). 1911.05722.

[29] Alzubaidi, L. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data* **8**, 53 (2021).

[30] LeCun, Y. & Bengio, Y. *Convolutional Networks for Images, Speech, and Time Series*, 255–258 (MIT Press, Cambridge, MA, USA, 1998).

[31] Luo, W., Li, Y., Urtasun, R. & Zemel, R. Understanding the effective receptive field in deep convolutional neural networks (2017). 1701.04128.

[32] Olah, C., Mordvintsev, A. & Schubert, L. Feature visualization. *Distill* (2017). Https://distill.pub/2017/feature-visualization.

[33] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition (2015). 1512.03385.

[34] Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks (2017). 1611.05431.

[35] Hu, J., Shen, L., Albanie, S., Sun, G. & Wu, E. Squeeze-and-excitation networks (2019). 1709.01507.

[36] Li, X., Wang, W., Hu, X. & Yang, J. Selective kernel networks (2019). 1903.06586.

[37] Badrinarayanan, V., Kendall, A. & Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**, 2481–2495 (2017).

[38] Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation (2015). 1411.4038.

[39] Peng, C., Zhang, X., Yu, G., Luo, G. & Sun, J. Large kernel matters – improve semantic segmentation by global convolutional network (2017). 1703.02719.

[40] Dong, C., Loy, C. C., He, K. & Tang, X. Learning a deep convolutional network for image super-resolution. In Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T. (eds.) *Computer Vision – ECCV 2014*, 184–199 (Springer International Publishing, Cham, 2014).

[41] Dong, C., Loy, C. C., He, K. & Tang, X. Image super-resolution using deep convolutional networks (2015). 1501.00092.

[42] Singh, V. K. *et al.* Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network (2018). 1809.01687.

[43] Havaei, M. *et al.* Brain tumor segmentation with deep neural networks. *Medical Image Analysis* **35**, 18–31 (2017).

[44] Bakas, S. *et al.* Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge (2019). 1811.02629.

[45] Myronenko, A. 3d mri brain tumor segmentation using autoencoder regularization (2018). 1810.11654.

[46] Isensee, F., Kickingereder, P., Wick, W., Bendszus, M. & Maier-Hein, K. H. No new-net (2019). 1809.10483.

[47] Chang, K. *et al.* Residual convolutional neural network for the determination of IDH status in low- and High-Grade gliomas from MR imaging. *Clin Cancer Res* **24**, 1073–1081 (2017).

[48] Ertosun, M. G. & Rubin, D. L. Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks. *AMIA Annu Symp Proc* **2015**, 1899–1908 (2015).

[49] Neyshabur, B., Bhojanapalli, S., McAllester, D. & Srebro, N. Exploring generalization in deep learning (2017). 1706.08947.

[50] Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical Image Analysis* **42**, 60–88 (2017).

[51] Weese, J. & Lorenz, C. Four challenges in medical image analysis from an industrial perspective. *Medical Image Analysis* **33**, 44–49 (2016). URL https://www.sciencedirect.com/science/article/pii/S1361841516300998. 20th anniversary of the Medical Image Analysis journal (MedIA).

[52] de Bruijne, M. Machine learning approaches in medical image analysis: From detection to diagnosis. *Med Image Anal* **33**, 94–97 (2016).

[53] Guss, W. H. & Salakhutdinov, R. On characterizing the capacity of neural networks using algebraic topology (2018). 1802.04443.

[54] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

[55] Tompson, J., Goroshin, R., Jain, A., LeCun, Y. & Bregler, C. Efficient object localization using convolutional networks (2015). 1411.4280.

[56] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization (2019). 1711.05101.

[57] Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y. & Girshick, R. Detectron2. https://github.com/facebookresearch/detectron2 (2019).

[58] Simard, P., Steinkraus, D. & Platt, J. Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, 958–963 (2003).

[59] Lopes, R. G., Yin, D., Poole, B., Gilmer, J. & Cubuk, E. D. Improving robustness without sacrificing accuracy with patch gaussian augmentation (2019). 1906.02611.

[60] Lee, H., Hwang, S. J. & Shin, J. Self-supervised label augmentation via input transformations (2020). 1910.05872.

[61] Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V. & Le, Q. V. Autoaugment: Learning augmentation policies from data (2019). 1805.09501.

[62] Cubuk, E. D., Zoph, B., Shlens, J. & Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space (2019). 1909.13719.

[63] Hussain, Z., Gimenez, F., Yi, D. & Rubin, D. Differential data augmentation techniques for medical imaging classification tasks. *AMIA Annu Symp Proc* **2017**, 979–984 (2018).

[64] Yang, S. *et al.* Image data augmentation for deep learning: A survey (2022). 2204.08610.

[65] Safdar, M. F., Alkobaisi, S. S. & Zahra, F. T. A comparative analysis of data augmentation approaches for magnetic resonance imaging (MRI) scan images of brain tumor. *Acta Inform Med* **28**, 29–36 (2020).

[66] Eaton-Rosen, Z., Bragman, F. J. S., Ourselin, S. & Cardoso, M. J. Improving data augmentation for medical image segmentation (2018).

[67] He, K., Girshick, R. & Dollár, P. Rethinking imagenet pre-training (2018). 1811.08883.

[68] Chen, T., Kornblith, S., Swersky, K., Norouzi, M. & Hinton, G. Big self-supervised models are strong semi-supervised learners (2020). 2006.10029.

[69] Zoph, B. *et al.* Rethinking pre-training and self-training (2020). 2006.06882.

[70] Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation (2014). 1311.2524.

[71] Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302 (1945). URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1932409. https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.2307/1932409.

[72] Zhang, J., Shen, X., Zhuo, T. & Zhou, H. Brain tumor segmentation based on refined fully convolutional neural networks with a hierarchical dice loss (2018). 1712.09093.

[73] Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection (2018). 1708.02002.

[74] Kervadec, H. *et al.* Boundary loss for highly unbalanced segmentation. *Medical Image Analysis* **67**, 101851 (2021).

[75] He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn (2018). 1703.06870.

[76] Shrivastava, A., Gupta, A. & Girshick, R. Training region-based object detectors with online hard example mining (2016). 1604.03540.

[77] Wang, X., Zhang, R., Shen, C., Kong, T. & Li, L. Solo: A simple framework for instance segmentation (2021). 2106.15947.

[78] Li, L., Zhou, T., Wang, W., Li, J. & Yang, Y. Deep hierarchical semantic segmentation (2022). 2203.14335.

[79] Otsu, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9**, 62–66 (1979).

[80] Bradley, D. & Roth, G. Adaptive thresholding using the integral image. *Journal of Graphics Tools* **12**, 13–21 (2007). URL https://doi.org/10.1080/2151237X.2007.10129236. https://doi.org/10.1080/2151237X.2007.10129236.

[81] Mittal, H. *et al.* A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets. *Multimedia Tools and Applications* **81**, 35001–35026 (2022). URL https://doi.org/10.1007/s11042-021-10594-9.

[82] Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 1–22 (2018). URL https://doi.org/10.1111/j.2517-6161.1977.tb01600.x. https://academic.oup.com/jrsssb/article-pdf/39/1/1/49117094/jrsssb_39_1_1.pdf.

[83] Held, K. *et al.* Markov random field segmentation of brain mr images. *IEEE Transactions on Medical Imaging* **16**, 878–886 (1997).

[84] Kass, M., Witkin, A. & Terzopoulos, D. Snakes: Active contour models. *International Journal of Computer Vision* **1**, 321–331 (1988). URL `https://doi.org/10.1007/BF00133570`.

[85] Boykov, Y. & Jolly, M.-P. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 1, 105–112 vol.1 (2001).

[86] Jensen, P. M., Jeppesen, N., Dahl, A. B. & Dahl, V. A. Review of serial and parallel min-cut/max-flow algorithms for computer vision (2022). `2202.00418`.

[87] Avants, B. B., Tustison, N., Song, G. *et al.* Advanced normalization tools (ants). *Insight j* **2**, 1–35 (2009).

[88] Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A. & Gee, J. C. An Open Source Multivariate Framework for n-Tissue Segmentation with Evaluation on Public Data. *Neuroinformatics* **9**, 381–400 (2011). URL `https://doi.org/10.1007/s12021-011-9109-y`.

[89] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs (2017). `1606.00915`.

[90] Azad, R. *et al.* Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation (2022). `2208.00713`.

[91] Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V. & Babenko, A. Label-efficient semantic segmentation with diffusion models (2022). `2112.03126`.

[92] Strudel, R., Garcia, R., Laptev, I. & Schmid, C. Segmenter: Transformer for semantic segmentation (2021). `2105.05633`.

[93] He, Y. *et al.* Deep learning based 3d segmentation: A survey (2021). `2103.05423`.

[94] Baid, U. *et al.* The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification (2021). `2107.02314`.

[95] Menze, B. H. *et al.* The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging* **34**, 1993–2024 (2014).

[96] Bakas, S. *et al.* Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data* **4**, 170117 (2017).

[97] Bakas, S. *et al.* Segmentation Labels for the Pre-operative Scans of the TCGA-GBM collection [Data set]. *The Cancer Imaging Archive* (2017).

[98] Bakas, S. *et al.* Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection [Data Set]. *The Cancer Imaging Archive* (2017).

[99] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation (2015). 1505.04597.

[100] Patel, J. *et al.* Segmentation, survival prediction, and uncertainty estimation of gliomas from multimodal 3d mri using selective kernel networks. In Crimi, A. & Bakas, S. (eds.) *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 228–240 (Springer International Publishing, Cham, 2021).

[101] Isensee, F., Jaeger, P. F., Full, P. M., Vollmuth, P. & Maier-Hein, K. H. nnu-net for brain tumor segmentation (2020). 2011.00848.

[102] Futrega, M., Milesi, A., Marcinkiewicz, M. & Ribalta, P. Optimized u-net for brain tumor segmentation (2021). 2110.03352.

[103] Jiang, Z., Ding, C., Liu, M. & Tao, D. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In Crimi, A. & Bakas, S. (eds.) *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 231–241 (Springer International Publishing, Cham, 2020).

[104] Zhao, Y., Zhang, Y.-M. & Liu, C.-L. *Bag of Tricks for 3D MRI Brain Tumor Segmentation*, 210–220 (2020).

[105] Siddiquee, M. M. R. & Myronenko, A. Redundancy reduction in semantic segmentation of 3d brain tumor mris (2021). 2111.00742.

[106] Azad, R. *et al.* Medical image segmentation review: The success of u-net (2022). 2211.14830.

[107] Nag, S. Image registration techniques: A survey (2017). URL https://doi.org/10.31224/osf.io/rv65c.

[108] Berger, M. *Geometry i* (Springer Science & Business Media, 2009).

[109] Maintz, J. B. A. & Viergever. An overview of medical image registration methods (1998).

[110] Mok, T. C. W. & Chung, A. C. S. Affine medical image registration with coarse-to-fine vision transformer (2022). 2203.15216.

[111] Oliveira, F. P. M. & Tavares, J. M. R. S. Medical image registration: a review. *Comput Methods Biomech Biomed Engin* **17**, 73–93 (2012).

[112] Kuang, D. On reducing negative jacobian determinant of the deformation predicted by deep registration networks (2019). 1907.00068.

[113] Zou, J., Gao, B., Song, Y. & Qin, J. A review of deep learning-based deformable medical image registration. *Frontiers in Oncology* **12** (2022). URL `https://www.frontiersin.org/articles/10.3389/fonc.2022.1047215`.

[114] Adelson, E. H., Burt, P. J., Anderson, C. H., Ogden, J. M. & Bergen, J. R. Pyramid methods in image processing. (1984).

[115] Rueckert, D. *et al.* Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans Med Imaging* **18**, 712–721 (1999).

[116] Thirion, J.-P. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Analysis* **2**, 243–260 (1998). URL `https://www.sciencedirect.com/science/article/pii/S1361841598800224`.

[117] Glaunes, J., Qiu, A., Miller, M. I. & Younes, L. Large deformation diffeomorphic metric curve mapping. *Int J Comput Vis* **80**, 317–336 (2008).

[118] Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal* **12**, 26–41 (2007).

[119] Avants, B. B. *et al.* A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* **54**, 2033–2044 (2010).

[120] Klein, A. *et al.* Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *NeuroImage* **46**, 786–802 (2009). URL `https://www.sciencedirect.com/science/article/pii/S1053811908012974`.

[121] Chee, E. & Wu, Z. Airnet: Self-supervised affine registration for 3d medical images using neural networks (2018). `1810.02583`.

[122] Islam, K. T., Wijewickrema, S. & O'Leary, S. A deep learning based framework for the registration of three dimensional multi-modal medical images of the head. *Scientific Reports* **11**, 1860 (2021).

[123] Sokooti, H. *et al.* Nonrigid image registration using multi-scale 3d convolutional neural networks. In Descoteaux, M. *et al.* (eds.) *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, 232–239 (Springer International Publishing, Cham, 2017).

[124] Dalca, A. V., Balakrishnan, G., Guttag, J. & Sabuncu, M. R. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical Image Analysis* **57**, 226–236 (2019).

[125] Mok, T. C. W. & Chung, A. C. S. Large deformation diffeomorphic image registration with laplacian pyramid networks (2020). `2006.16148`.

[126] de Vos, B. D. *et al.* A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis* **52**, 128–143 (2019).

[127] Christodoulidis, S. *et al.* Linear and deformable image registration with 3d convolutional neural networks (2018). 1809.06226.

[128] Grumbach, M. M. *et al.* Management of the clinically inapparent adrenal mass ("incidentaloma"). *Ann Intern Med* **138**, 424–429 (2003).

[129] Choyke, P. L. & ACR Committee on Appropriateness Criteria. ACR appropriateness criteria on incidentally discovered adrenal mass. *J Am Coll Radiol* **3**, 498–504 (2006).

[130] Mayo-Smith, W. W., Boland, G. W., Noto, R. B. & Lee, M. J. State-of-the-art adrenal imaging. *RadioGraphics* **21**, 995–1012 (2001). URL https://doi.org/10.1148/radiographics.21.4.g01jl21995. PMID: 11452074, https://doi.org/10.1148/radiographics.21.4.g01jl21995.

[131] Sahdev, A. & Reznek, R. H. The indeterminate adrenal mass in patients with cancer. *Cancer Imaging* **7 Spec No A**, S100–9 (2007).

[132] Mayo-Smith, W. W. *et al.* Management of incidental adrenal masses: A white paper of the ACR incidental findings committee. *J Am Coll Radiol* **14**, 1038–1044 (2017).

[133] Koyuncu, H., Ceylan, R., Erdogan, H. & Sivri, M. A novel pipeline for adrenal tumour segmentation. *Computer Methods and Programs in Biomedicine* **159**, 77–86 (2018). URL https://www.sciencedirect.com/science/article/pii/S0169260717308295.

[134] Tabouret, E. *et al.* Recent trends in epidemiology of brain metastases: an overview. *Anticancer Res* **32**, 4655–4662 (2012).

[135] Langer, C. J. & Mehta, M. P. Current management of brain metastases, with a focus on systemic options. *J Clin Oncol* **23**, 6207–6219 (2005).

[136] Lin, N. U. *et al.* Response assessment criteria for brain metastases: proposal from the RANO group. *Lancet Oncol* **16**, e270–8 (2015).

[137] Soffietti, R., Chiavazza, C. & Rudà, R. Imaging and clinical end points in brain metastases trials. *CNS Oncol* **6**, 243–246 (2017).

[138] Bauknecht, H.-C. *et al.* Intra- and interobserver variability of linear and volumetric measurements of brain metastases using contrast-enhanced magnetic resonance imaging. *Invest Radiol* **45**, 49–56 (2010).

[139] Bousabarah, K. *et al.* Deep convolutional neural networks for automated segmentation of brain metastases trained on clinical data. *Radiat Oncol* **15**, 87 (2020).

[140] Grøvik, E. *et al.* Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. *J Magn Reson Imaging* **51**, 175–182 (2019).

[141] Young, W. F. The incidentally discovered adrenal mass. *New England Journal of Medicine* **356**, 601–610 (2007). URL `https://doi.org/10.1056/NEJMcp065470`. PMID: 17287480, `https://doi.org/10.1056/NEJMcp065470`.

[142] Glazer, D. I., Corwin, M. T. & Mayo-Smith, W. W. Incidental adrenal nodules. *Radiol Clin North Am* **59**, 591–601 (2021).

[143] Song, J. H., Chaudhry, F. S. & Mayo-Smith, W. W. The incidental adrenal mass on ct: Prevalence of adrenal disease in 1,049 consecutive adrenal masses in patients with no known malignancy. *American Journal of Roentgenology* **190**, 1163–1168 (2008). URL `https://doi.org/10.2214/AJR.07.2799`. PMID: 18430826, `https://doi.org/10.2214/AJR.07.2799`.

[144] Berland, L. L. *et al.* Managing incidental findings on abdominal CT: white paper of the ACR incidental findings committee. *J Am Coll Radiol* **7**, 754–773 (2010).

[145] Zeiger, M. A. *et al.* The american association of clinical endocrinologists and american association of endocrine surgeons medical guidelines for the management of adrenal incidentalomas. *Endocr Pract* **15 Suppl 1**, 1–20 (2009).

[146] Kebebew, E. Adrenal incidentaloma. *N Engl J Med* **384**, 1542–1551 (2021).

[147] Wang, S. & Summers, R. M. Machine learning and radiology. *Med Image Anal* **16**, 933–951 (2012).

[148] Kavur, A. E. *et al.* CHAOS challenge - combined (CT-MR) healthy abdominal organ segmentation. *Med Image Anal* **69**, 101950 (2020).

[149] Tong, N., Gou, S., Niu, T., Yang, S. & Sheng, K. Self-paced DenseNet with boundary constraint for automated multi-organ segmentation on abdominal CT images. *Phys Med Biol* **65**, 135011 (2020).

[150] Jimenez-Pastor, A. *et al.* Precise whole liver automatic segmentation and quantification of PDFF and r2* on MR images. *Eur Radiol* **31**, 7876–7887 (2021).

[151] Nelms, B. E., Tomé, W. A., Robinson, G. & Wheeler, J. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. *Int J Radiat Oncol Biol Phys* **82**, 368–378 (2010).

[152] Kim, H. *et al.* Abdominal multi-organ auto-segmentation using 3D-patch-based deep convolutional neural network. *Scientific Reports* **10**, 6204 (2020). URL `https://doi.org/10.1038/s41598-020-63285-0`.

[153] Gibson, E. *et al.* Automatic Multi-Organ segmentation on abdominal CT with dense V-Networks. *IEEE Trans Med Imaging* **37**, 1822–1834 (2018).

[154] Koyuncu, H., Ceylan, R., Erdogan, H. & Sivri, M. A novel pipeline for adrenal tumour segmentation. *Comput Methods Programs Biomed* **159**, 77–86 (2018).

[155] Saiprasad, G., Chang, C.-I., Safdar, N., Saenz, N. & Siegel, E. Adrenal gland abnormality detection using random forest classification. *J Digit Imaging* **26**, 891–897 (2013).

[156] Yi, X. *et al.* Adrenal incidentaloma: machine learning-based quantitative texture analysis of unenhanced CT can effectively differentiate sPHEO from lipid-poor adrenal adenoma. *J Cancer* **9**, 3577–3582 (2018).

[157] Romeo, V. *et al.* Characterization of adrenal lesions on unenhanced mri using texture analysis: A machine-learning approach. *Journal of Magnetic Resonance Imaging* **48**, 198–204 (2018). URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.25954`. `https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.25954`.

[158] Magudia, K. *et al.* Population-scale ct-based body composition analysis of a large outpatient population using deep learning to derive age-, sex-, and race-specific reference curves. *Radiology* **298**, 319–329 (2021). URL `https://doi.org/10.1148/radiol.2020201640`. PMID: 33231527, `https://doi.org/10.1148/radiol.2020201640`.

[159] Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks (2018). `1608.06993`.

[160] Lee, H., Kim, M. & Do, S. Practical window setting optimization for medical image deep learning (2018). `1812.00572`.

[161] Masoudi, S. *et al.* Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. *Journal of Medical Imaging* **8**, 010901 (2021). URL `https://doi.org/10.1117/1.JMI.8.1.010901`.

[162] Wu, Y. & He, K. Group normalization (2018). `1803.08494`.

[163] Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, 807–814 (Omnipress, Madison, WI, USA, 2010).

[164] Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift (2015). `1502.03167`.

[165] Kayalibay, B., Jensen, G. & van der Smagt, P. Cnn-based segmentation of medical imaging data (2017). `1701.03056`.

[166] Nagi, J. *et al.* Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 342–347 (2011).

[167] Strutz, T. The distance transform and its computation (2023). 2106.03503.

[168] McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* **22**, 276–282 (2012).

[169] Van Rossum, G. & Drake Jr, F. L. *Python tutorial* (Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995).

[170] "july effect": Impact of the academic year-end changeover on patient outcomes. *Annals of Internal Medicine* **155**, 309–315 (2011). URL https://www.acpjournals.org/doi/abs/10.7326/0003-4819-155-5-201109060-00354. PMID: 21747093, https://www.acpjournals.org/doi/pdf/10.7326/0003-4819-155-5-201109060-00354.

[171] Minnaar, E. M. *et al.* An adrenal incidentaloma: how often is it detected and what are the consequences? *ISRN Radiol* **2013**, 871959 (2012).

[172] Barstuğan, M., Ceylan, R., Asoglu, S., Cebeci, H. & Koplay, M. Adrenal tumor segmentation method for MR images. *Comput Methods Programs Biomed* **164**, 87–100 (2018).

[173] Chai, H., Guo, Y., Wang, Y. & Zhou, G. Automatic computer aided analysis algorithms and system for adrenal tumors on CT images. *Technol Health Care* **25**, 1105–1118 (2017).

[174] Brain metastases. *Nature Reviews Disease Primers* **5**, 6 (2019).

[175] Nayak, L., Lee, E. Q. & Wen, P. Y. Epidemiology of brain metastases. *Current oncology reports* **14**, 48–54 (2012).

[176] Lignelli, A. & Khandji, A. G. Review of imaging techniques in the diagnosis and management of brain metastases. *Neurosurgery Clinics* **22**, 15–25 (2011).

[177] Fabi, A. *et al.* Brain metastases from solid tumors: disease outcome according to type of treatment and therapeutic resources of the treating center. *J Exp Clin Cancer Res* **30**, 10 (2011).

[178] Tong, E., McCullagh, K. L. & Iv, M. Advanced imaging of brain metastases: from augmenting visualization and improving diagnosis to evaluating treatment response. *Frontiers in Neurology* **11**, 270 (2020).

[179] Lin, N. U. *et al.* Response assessment criteria for brain metastases: proposal from the RANO group. *Lancet Oncol* **16**, e270–8 (2015).

[180] Growcott, S., Dembrey, T., Patel, R., Eaton, D. & Cameron, A. Inter-observer variability in target volume delineations of benign and metastatic brain tumours for stereotactic radiosurgery: Results of a national quality assurance programme. *Clinical Oncology* **32**, 13–25 (2020). URL https://www.sciencedirect.com/science/article/pii/S0936655519302766.

[181] Ellingson, B. M. *et al.* Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials. *Neuro Oncol* **17**, 1188–1198 (2015).

[182] Wolpert, F. *et al.* Risk factors for the development of epilepsy in patients with brain metastases. *Neuro-oncology* **22**, 718–728 (2020).

[183] Nussbaum, E. S., Djalilian, H. R., Cho, K. H. & Hall, W. A. Brain metastases. histology, multiplicity, surgery, and survival. *Cancer* **78**, 1781–1788 (1996).

[184] Wang, S. *et al.* Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Medical Image Analysis* (2017).

[185] Havaei, M. *et al.* Brain tumor segmentation with Deep Neural Networks. *Medical Image Analysis* **35**, 18–31 (2017).

[186] Li, X. *et al.* H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes. *IEEE Transactions on Medical Imaging* (2018).

[187] Chang, K. *et al.* Automatic assessment of glioma burden: A deep learning algorithm for fully automated volumetric and bidimensional measurement. *Neuro-Oncology* (2019).

[188] Kickingereder, P. *et al.* Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol* **20**, 728–740 (2019).

[189] Ottesen, J. A. *et al.* 2.5d and 3D segmentation of brain metastases with deep learning on multinational MRI data. *Front Neuroinform* **16**, 1056068 (2023).

[190] Rudie, J. D. *et al.* Three-dimensional u-net convolutional neural network for detection and segmentation of intracranial metastases. *Radiology: Artificial Intelligence* **3**, e200204 (2021). URL `https://doi.org/10.1148/ryai.2021200204`. `https://doi.org/10.1148/ryai.2021200204`.

[191] Dikici, E. *et al.* Automated brain metastases detection framework for t1-weighted contrast-enhanced 3d mri. *IEEE journal of biomedical and health informatics* **24**, 2883–2893 (2020).

[192] Cho, S. J. *et al.* Brain metastasis detection using machine learning: a systematic review and meta-analysis. *Neuro-oncology* **23**, 214–225 (2021).

[193] Bousabarah, K. *et al.* Deep convolutional neural networks for automated segmentation of brain metastases trained on clinical data. *Radiation Oncology* **15**, 1–9 (2020).

[194] Xue, J. *et al.* Deep learning–based detection and segmentation-assisted management of brain metastases. *Neuro-oncology* **22**, 505–514 (2020).

[195] Brastianos, P. K. *et al.* Single-arm, open-label phase 2 trial of pembrolizumab in patients with leptomeningeal carcinomatosis. *Nature medicine* **26**, 1280–1284 (2020).

[196] Leone, J. P. *et al.* A phase II study of cabozantinib alone or in combination with trastuzumab in breast cancer patients with brain metastases. *Breast cancer research and treatment* **179**, 113–123 (2020).

[197] Fedorov, A. *et al.* 3D slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging* **30**, 1323–1341 (2012).

[198] Speier, W., Iglesias, J. E., El-Kara, L., Tu, Z. & Arnold, C. Robust skull stripping of clinical glioblastoma multiforme data. *Med Image Comput Comput Assist Interv* **14**, 659–666 (2011).

[199] Tustison, N. J. *et al.* N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging* **29**, 1310–1320 (2010).

[200] Ulyanov, D., Vedaldi, A. & Lempitsky, V. Instance normalization: The missing ingredient for fast stylization (2017). `1607.08022`.

[201] Guerrero-Pena, F. A. *et al.* Multiclass Weighted Loss for Instance Segmentation of Cluttered Cells. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (IEEE, 2018).

[202] Hoebel, K. *et al.* An exploration of uncertainty information for segmentation quality assessment. In Išgum, I. & Landman, B. A. (eds.) *Medical Imaging 2020: Image Processing*, vol. 11313, 113131K. International Society for Optics and Photonics (SPIE, 2020). URL `https://doi.org/10.1117/12.2548722`.

[203] Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* **15**, 155–163 (2016).

[204] Peng, J. *et al.* Deep learning-based automatic tumor burden assessment of pediatric high-grade gliomas, medulloblastomas, and other leptomeningeal seeding tumors. *Neuro-oncology* **24**, 289–299 (2022).

[205] Sermesant, M., Delingette, H., Cochet, H., Jaïs, P. & Ayache, N. Applications of artificial intelligence in cardiovascular imaging. *Nature Reviews Cardiology* **18**, 600–609 (2021). URL `https://doi.org/10.1038/s41569-021-00527-2`.

[206] Durkee, M. S., Abraham, R., Clark, M. R. & Giger, M. L. Artificial intelligence and cellular segmentation in tissue microscopy images. *The American Journal of Pathology* **191**, 1693–1701 (2021). URL `https://www.sciencedirect.com/science/article/pii/S0002944021002613`.

[207] Chen, W., Wang, Y., Tian, D. & Yao, Y. Ct lung nodule segmentation: A comparative study of data preprocessing and deep learning models. *IEEE Access* **11**, 34925–34931 (2023).

[208] Robinson-Weiss, C. *et al.* Machine learning for adrenal gland segmentation and classification of normal and adrenal masses at CT. *Radiology* **306**, e220101 (2022).

[209] Tabouret, E. *et al.* Recent trends in epidemiology of brain metastases: an overview. *Anticancer Res* **32**, 4655–4662 (2012).

[210] Sperduto, P. W. *et al.* Survival in patients with brain metastases: Summary report on the updated diagnosis-specific graded prognostic assessment and definition of the eligibility quotient. *Journal of Clinical Oncology* **38**, 3773–3784 (2020). URL `https://doi.org/10.1200/JCO.20.01255`. PMID: 32931399, `https://doi.org/10.1200/JCO.20.01255`.

[211] Li, A. Y. *et al.* Association of Brain Metastases With Survival in Patients With Limited or Stable Extracranial Disease: A Systematic Review and Meta-analysis. *JAMA Network Open* **6**, e230475–e230475 (2023). URL `https://doi.org/10.1001/jamanetworkopen.2023.0475`. `https://jamanetwork.com/journals/jamanetworkopen/articlepdf/2801743/li_2023_oi_230031_1676399886.26225.pdf`.

[212] Cagney, D. N. *et al.* Incidence and prognosis of patients with brain metastases at diagnosis of systemic malignancy: a population-based study. *Neuro Oncol* **19**, 1511–1521 (2017).

[213] Vogelbaum, M. A. *et al.* Treatment for brain metastases: Asco-sno-astro guideline. *Journal of Clinical Oncology* **40**, 492–516 (2022). URL `https://doi.org/10.1200/JCO.21.02314`. PMID: 34932393, `https://doi.org/10.1200/JCO.21.02314`.

[214] Menze, B. H. *et al.* The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging* **34**, 1993–2024 (2014).

[215] Bakas, S. *et al.* Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data* **4**, 170117 (2017).

[216] Cheng, V. W. *et al.* VCAM-1–targeted MRI Enables Detection of Brain Micrometastases from Different Primary Tumors. *Clinical Cancer Research* **25**, 533–543 (2019). URL `https://doi.org/10.1158/1078-0432.CCR-18-1889`. `https://aacrjournals.org/clincancerres/article-pdf/25/2/533/2302316/533.pdf`.

[217] Nomoto, Y., Miyamoto, T. & Yamaguchi, Y. Brain metastasis of small cell lung carcinoma: comparison of Gd-DTPA enhanced magnetic resonance imaging and enhanced computerized tomography. *Jpn J Clin Oncol* **24**, 258–262 (1994).

[218] Grøvik, E. *et al.* Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. *J Magn Reson Imaging* **51**, 175–182 (2019).

[219] Qin, C. *et al.* Joint learning of motion estimation and segmentation for cardiac mr image sequences (2018). `1806.04066`.

[220] Upendra, R. R., Simon, R. & Linte, C. A. Joint deep learning framework for image registration and segmentation of late gadolinium enhanced MRI and cine cardiac MRI. *Proc SPIE Int Soc Opt Eng* **11598** (2021).

[221] Xu, Z. & Niethammer, M. Deepatlas: Joint semi-supervised learning of image registration and segmentation (2019). `1904.08465`.

[222] Chen, X., Xia, Y., Ravikumar, N. & Frangi, A. F. Joint segmentation and discontinuity-preserving deformable registration: Application to cardiac cine-mr images (2022). `2211.13828`.

[223] Jaderberg, M., Simonyan, K., Zisserman, A. & Kavukcuoglu, K. Spatial transformer networks (2016). `1506.02025`.

[224] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014). URL `http://jmlr.org/papers/v15/srivastava14a.html`.

[225] Wilcoxon, F. *Individual Comparisons by Ranking Methods*, 196–202 (Springer New York, New York, NY, 1992). URL `https://doi.org/10.1007/978-1-4612-4380-9_16`.

[226] Huttenlocher, D., Klanderman, G. & Rucklidge, W. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**, 850–863 (1993).

[227] Inc., T. M. Matlab version: 9.5.0 (r2018b) (2022). URL `https://www.mathworks.com`.

[228] Beers, A. *et al.* DeepNeuro: an open-source deep learning toolbox for neuroimaging. *Neuroinformatics* (2020). URL `https://doi.org/10.1007/s12021-020-09477-5`.

[229] Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems (2015). URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[230] Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization (2017). `1711.05101`.

[231] He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Computer Vision – ECCV 2014*, 346–361 (Springer International Publishing, 2014).

[232] Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R. & Horvath, P. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific Reports* **10**, 5068 (2020).

[233] Ashraf, H., Waris, A., Ghafoor, M. F., Gilani, S. O. & Niazi, I. K. Melanoma segmentation using deep learning with test-time augmentations and conditional random fields. *Scientific Reports* **12**, 3948 (2022).

[234] Lu, H., Shanmugam, D., Suresh, H. & Guttag, J. Improved text classification via test-time augmentation (2022). 2206.13607.

[235] Tawbi, H. A. *et al.* Combined nivolumab and ipilimumab in melanoma metastatic to the brain. *New England Journal of Medicine* **379**, 722–730 (2018). URL https://doi.org/10.1056/NEJMoa1805453. PMID: 30134131, https://doi.org/10.1056/NEJMoa1805453.

[236] Goldberg, S. B. *et al.* Pembrolizumab for management of patients with NSCLC and brain metastases: long-term results and biomarker analysis from a non-randomised, open-label, phase 2 trial. *Lancet Oncol* **21**, 655–663 (2020).

[237] Brastianos, P. K. *et al.* Palbociclib demonstrates intracranial activity in progressive brain metastases harboring cyclin-dependent kinase pathway alterations. *Nature Cancer* **2**, 498–502 (2021). URL https://doi.org/10.1038/s43018-021-00198-5.

[238] Petersen, J. *et al.* Deep probabilistic modeling of glioma growth (2019). 1907.04064.

[239] Jacobs, M., Kim, I. & Tong, J. Tumor growth with nutrients: Regularity and stability (2022). 2204.07572.

[240] Al-Huniti, N. *et al.* Tumor growth dynamic modeling in oncology drug development and regulatory approval: Past, present, and future opportunities. *CPT Pharmacometrics Syst Pharmacol* **9**, 419–427 (2020).

[241] Modat, M. *et al.* Fast free-form deformation using graphics processing units. *Computer Methods and Programs in Biomedicine* **98**, 278–284 (2010). URL https://www.sciencedirect.com/science/article/pii/S0169260709002533. HP-MICCAI 2008.

[242] Li, H., Fan, Y. & for the Alzheimer's Disease Neuroimaging Initiative. Mdreg-net: Multi-resolution diffeomorphic image registration using fully convolutional networks with deep self-supervision. *Human Brain Mapping* **43**, 2218–2231 (2022). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25782. https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.25782.

[243] Liu, L., Hu, X., Zhu, L. & Heng, P.-A. Probabilistic multilayer regularization network for unsupervised 3d brain image registration. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 346–354 (Springer, 2019).