Active Learning as an Approach to Deep Learning Glioma Segmentation from Brain MR Images

By Andrew Boehringer

A thesis submitted to the Faculty of Medicine to obtain the university interdisciplinary Master's degree in Neuroscience of the University of Geneva.

Master Supervisor:
Prof. Habib Zaidi

Jury:
Prof. Habib Zaidi
Prof. Domenico Della Volpe
Prof. Maria Isabel Vargas

Geneva, 01/07/2022

Acknowledgements

I would like to thank Professor Habib Zaidi for sharing his guidance and expertise with me and for all his support in my research and studies.

I would also like to thank Dr. Hossein Arabi and PhD student Amirhossein Sanaat for sharing their knowledge and suggestions with me and for their mentorship throughout the program.

I would like to thank all the members of PINLAB who were very welcoming and supportive to me and provided me with an excellent and collaborative working environment.

I am deeply grateful to my family and friends for their support and encouragement throughout the program.

Abstract

When training deep learning algorithms, a large amount of manually annotated data is required as standard of reference or ground truth. The manual annotation of these data can be both a laborious and time-consuming task, requiring the valuable time of trained experts. Therefore, an approach to reduce the amount of time needed to prepare enough ground truth data to adequately train a deep learning model is desirable. Active learning techniques aim to tackle this issue by decreasing the amount of ground truth data needed in model training using a semi-supervised approach in which the model is first trained with a smaller dataset, and then from this model the unlabeled cases which would most benefit the model are queried for manual annotation and added to the dataset. This approach would reduce the labor burden of manual annotation by using the most informative instances to train the model rather than random selection. This study focuses on assessing the application of active learning techniques to train a brain MRI glioma segmentation model. The publicly available training dataset provided for the 2021 RSNA-ASNR-MICCAI Brain Tumor Segmentation (BraTS) Challenge was used in this study, consisting of 1251 multi-institutional, multi-parametric MRI scans with pathologically confirmed glioma diagnosis. Post-contrast T1 (T1c), T2 (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR) images as well as ground truth manual segmentation performed by expert neuroradiologists were used as input for the model. The data were split into a training set of 1151 cases and testing set of 100 cases, with the testing set remaining constant throughout. Deep convolutional neural network segmentation models were trained using the NiftyNet platform. To test the viability of active learning in training a segmentation model, an initial reference model was trained using all 1151 training cases followed by two additional models using only 575 cases and 100 cases. The resulting predicted segmentations of these two additional models on the remaining training cases were then addended to the training dataset for additional training. For cases with a Dice score above a 0.7 threshold, the predicted segmentations were added in place of ground truth, while for those below the threshold the ground truth segmentations were used. The new updated training sets were then used to continue model training and the resulting models compared. In clinical settings, the ground truth data are not as readily available as in research settings and as such, it was important to also consider how the unlabeled data is queried for manual segmentation in the absence of ground truth data. For this purpose, a secondary classification model was trained using MATLAB's deep learning module to predict the segmentation quality when provided with no ground truth for comparison. In addition to T1c, T2, and FLAIR inputs, the segmentation probability maps provided by the segmentation models were included as input. Segmentation classes were "poor quality" (Dice score < 0.6), "acceptable with adjustments" (Dice score between 0.6 and 0.8), and "acceptable quality" (Dice score > 0.8). It was demonstrated that an active learning approach for manual segmentation can lead to comparable model performance for segmentation of brain gliomas at a lower manual annotation requirement. Additionally, secondary models can be developed to determine which cases the segmentation model will struggle with so that they can be used for most efficient model training. Though the performance of the segmentation models in this study would need to be improved upon to be applied in clinical practice, an active learning approach may benefit future

model training by streamlining the training process in problems that require large amounts of training data.

Keywords: MRI, brain imaging, gliomas, segmentation, deep learning.

Table of Contents

1 Introduction 1

1.1 Magnetic Resonance Imaging 1

1.2 Brain Gliomas 3

1.3 Medical Image Segmentation 4

1.4 Deep Learning in Medical Imaging 4

1.5 Active Learning 5

2 Materials and Methods 7

2.1 BraTS Dataset 7

2.2 Data Preprocessing 8

2.3 Training Reference and Baseline Models 9

2.4 Evaluation of Baseline Models 10

2.5 Active Learning 12

2.6 Secondary Classification Model 13

2.7 Training Secondary Classification Model 14

3 Results 17

3.1 Baseline Model Metric Results 17

3.2 Training Process 18

3.3 Active Learning Results 20

3.4 Classification Model Results 23

4 Discussion 26

5 Conclusion 29

References 30

List of Abbreviations

3D	3-Dimensional
AUC	Area under the receiver operating characteristic curve
BraTS	Brain tumor segmentation
ED	Peritumoral edematous/invaded tissue
ET	Enhancing Tumor
FLAIR	T2 Fluid attenuated inversion recovery
FN	False negative
FP	False positive
Gd	Gadolinium
MRI	Magnetic resonance imaging
NCR	Necrotic core
PPV	Positive predictive value
ReLU	Rectified linear unit
RF	Radiofrequency
ROC	Receiver operating characteristic
T1	Pre-contrast T1-weighted MRI
T100	Testing set of 100 cases
T1c	Post-Gadolinium contrast T1-weighted MRI
T2	Non-contrast T2-weighted MRI
TN	True negative
TP	True positive
WT	Whole Tumor

List of Figures

1.1. Representative T2 and T2-FLAIR MR images. 2

1.2. Representative T1-weighted MR images with and without contrast. 3

1.3. An example schematic demonstrating the active learning process. 6

2.1. Representative images of T1c, T2, FLAIR, and manual ground truth segmentation. 7

2.2. Union of NCR, ET, and ED labels into the single WT label. 8

2.3. Network architecture of the segmentation model. 9

2.4. Example input images for the classification model. 14

2.5. Network architecture of the classification model. 15

3.1. Training progress of the three baseline models. 19

3.2. Training progress of Model B's baseline and active learning models. 19

3.3. Training progress of Model C's baseline and active learning models. 20

3.4. Post-active learning predicted segmentations for a representative case. 22

3.5. Example cases with low and high Dice scores. 23

3.6. Confusion matrix for the classification model. 24

3.7. ROC curves for the classification model. 25

List of Tables

2.1. Training parameters for the glioma segmentation model. 10

2.2. Summary of the training cases used in each baseline model. 10

2.3. Summary of the training cases used in each active learning model. 13

2.4. Training parameters for the classification model. 15

3.1. Metric results for each baseline model. 17

3.2. Metric results for each model iteration selected for active learning. 18

3.3. Metric results of the pre- and post-active learning segmentation models. 21

3.4. Metric results for each class of the classification model. 25

Chapter 1

Introduction

1.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a medical imaging technique used to view anatomical structures within the body. The benefits of MRI over other imaging techniques are the high image resolution obtained as well as the use of non-ionizing radiation. MRI utilizes strong magnets, often at strengths of either 1.5 or 3.0 Tesla, as well as radiofrequency (RF) pulses to measure the properties of excited hydrogen atoms back to their equilibrium state. Hydrogen is used due to its presence in both water and lipids in making up 75-80% of the human body (McRobbie, 2007). Through these measurements, the hydrogen composition at each point can be inferred and used to generate an image. The key measurements made that affect the contrast of the image are the proton density, spin-lattice relaxation time (or T1), and spin-spin relaxation time (or T2) (McRobbie, 2007).

MRI acquisition utilizes three different types of magnetic fields: the static magnetic field of the scanner, the oscillating magnetic fields of the RF pulse, and gradients used for spatial localization (McRobbie, 2007). In the static magnetic field introduced by the machine, the hydrogen nuclei are all aligned parallel with the magnetic field. Using the RF pulse, the nuclei are flipped 90° into the transverse plane and begin to precess back to the equilibrium of the static magnetic field at their Larmor frequency shown in Equation 1.1, for which γ is the gyromagnetic ratio and β_0 is the strength of the magnetic field (McRobbie, 2007). From here, echoes are created to enhance the amplitude of the signal for easier measurement. Echoes can be either gradient echoes or spin echoes. In gradient echoes, a negative gradient is applied immediately after the RF pulse and followed by a positive gradient. This reverses the magnetic field gradient so that low frequency spin precessions now precess at higher frequencies due to gradient position (McRobbie, 2007). In spin echoes, after the 90° RF pulse, a second 180° RF pulse is applied to flip the spins. Though the precessional frequencies do not change, the phase angles are reversed and after a time equal to the delay between the two RF pulses, the spins come back into phase forming the echo (McRobbie, 2007).

$$\omega_0 = \gamma\beta_0 \tag{1.1}$$

By manipulating various parameters during MRI acquisition, different sequences can be acquired to achieve a desired image appearance. Some example sequences include T1-weighted, T2-weighted, and T2 Fluid Attenuated Inversion Recovery (FLAIR). In T1-weighted imaging a short repetition time and short echo time are used, resulting in dark fluids, bright fats, and with water-based tissues in between. In T2-weighted imaging a long repetition time and long echo time are used, resulting in images with bright fluids while fats and water-based

tissues are grey. T2-FLAIR sequences are a modification of the T2-weighted sequence that suppresses signal from cerebrospinal fluid and can be very useful in viewing brain lesions (Bakshi et al., 2001; De Coene et al., 1992). T2-FLAIR demonstrates well how the adjustments during the acquisition can have beneficial results on the output image appearance. The T2-weighted FLAIR sequence utilizes spin echoes with a 180° inversion recovery pulse and with the inversion type specifically selected so that recovery of most of the brain magnetization occurs while greatly reducing the artifacts from cerebrospinal fluid (De Coene et al., 1992). Representative T2 and T2-FLAIR images can be seen in Figure 1.1.

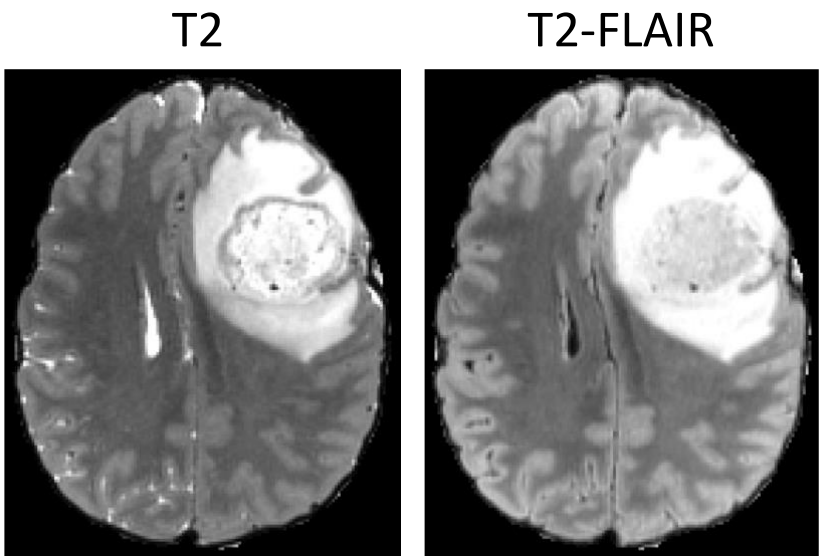

Figure 1.1. Representative T2-weighted (left) and T2-FLAIR (right) MR images demonstrating the reduction in contrast of the cerebrospinal fluid in the T2-FLAIR sequence.

In addition to varying the sequence to adjust image contrasts, chemical contrast agents, such as Gadolinium (Gd), can also be injected into the patient that show brightly in the resulting image and can be used to highlight desired structures. An example of the effect that contrast agents have on resulting MR images can be seen in Figure 1.2. Contrast agents work by shortening the relaxation time in tissues with contrast uptake. These are often used in detection of abnormal tissue structures such as tumors.

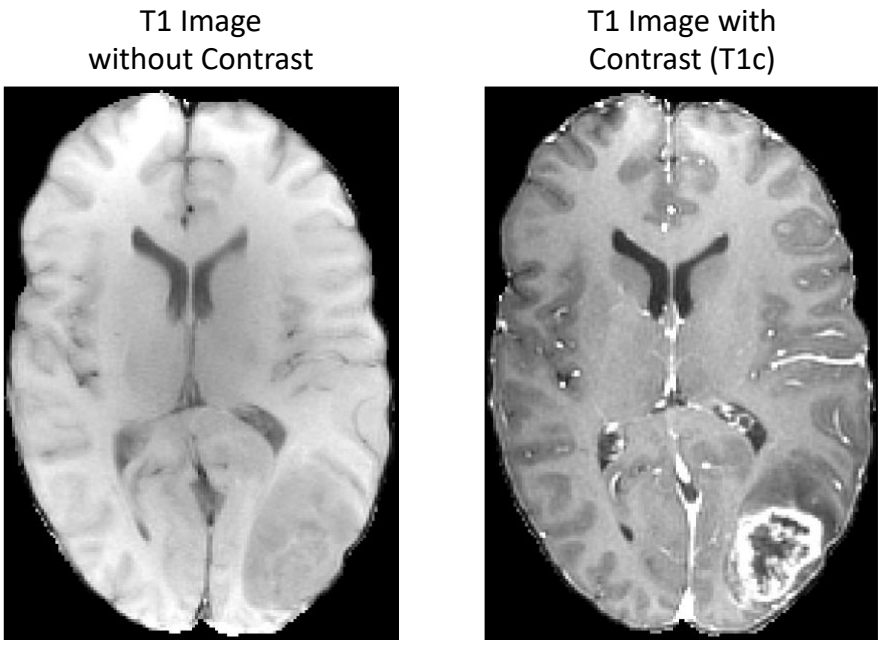

Figure 1.2. Representative T1-weighted images before (left) and after (right) the injection of a gadolinium contrast agent.

1.2 Brain Gliomas

Brain tumors are a subset of tumors that have a disproportionately high number of cancer-related deaths despite making up only 2% of all cancer incidences (Neugut et al., 2019). Primary brain tumors can vary in morphology and fall under many subtypes. Among these some examples include, in order of prevalence: glioblastoma, meningioma, other astrocytic tumors (excluding glioblastoma), oligodendroglial tumors, primary central nervous system lymphoma, ependymal tumors, oligoastrocytoma, and embryonal tumors (Wanis et al., 2021). Among patients diagnosed with malignant brain tumors, the 5-year survival rate is about 34.4%, and the most common type of malignant brain tumor, glioblastoma, has a 5-year survival rate of only about 5% (Ostrom et al., 2020) and so screening for early detection and treatment can be critical in improving outcomes (Croswell et al., 2010).

A glioma is a category of brain tumor forming in the glial cells of the central nervous system. They are the most common primary malignant brain tumor in adults and make up roughly 80% of malignant brain tumors (Chen et al., 2017; Ostrom et al., 2014). As of 2021, the WHO Classification of Tumors of the Central Nervous System (WHO Classification of Tumours Editorial Board, 2021) classifies gliomas in 6 subtypes: adult-type diffuse gliomas, pediatric-type diffuse low-grade gliomas, pediatric-type diffuse high-grade gliomas, circumscribed astrocytic gliomas, glioneural and neuronal tumors, and ependymomas (Louis et al., 2021). It also divides the most common types of adult gliomas into 3 sub-types: astrocytoma, oligodendroglioma, and glioblastoma (Louis et al., 2021). As mentioned before, glioblastomas, the most common type of glioma making up roughly 45% of gliomas, only have a 5-year survival rate of around 5% (Ostrom et al., 2014; Ostrom et al., 2020). Because of this, detection of gliomas is highly important to better the prognosis of the patients and improve

outcomes by determining the subtype as soon as possible for proper treatment. MRI is currently the most used technique for detection of gliomas in the clinical setting (Gokila Brindha et al., 2021; Wang et al., 2014).

1.3 Medical Image Segmentation

In imaging, segmentation is the process of partitioning an image into regions to represent meaningful areas for easier analysis (Shapiro & Stockman, 2001). Each of the pixels, or voxels in 3-dimensional imaging, of the partitioned segments share characteristics or features that allow them to be grouped (Nock & Nielsen, 2004). Each pixel/voxel in a segmented region is given a binary label of either being part of the segment or not, which are then combined to form a binary segmentation mask. In medical imaging, segmentation (Pham et al., 2000) is often used to delineate structures including organs (Fu et al., 2021), regions of the brain (Balafar et al., 2010), and tumors such as gliomas (Bauer et al., 2013; Wu et al., 2014). Segmentation of gliomas can be a very important step in both diagnosis and treatment planning. When diagnosing gliomas and planning for radiotherapy in cases of malignancy, for example, accurate segmentation can be useful in tasks such as determination of WHO grade, target volume and dosage planning, prediction of the location of tumor reoccurrence, and differentiation of pseudoprogression from actual tumor progression (Kocher et al., 2020; Mazzara et al., 2004). It is therefore crucial for segmentations to be as accurate as possible so that none of the tumor is missed and so that it can be properly diagnosed to receive the correct treatment approach.

1.4 Deep Learning in Medical Imaging

Neural networks are artificial networks of connected nodes used for decision-based problems and inspired by the communication of neurons in the human brain. Each connection represents a synapse between two nodes to pass on information. The connections, often referred to as edges, typically carry a weight that influences the strength of the signal passed on by the connection and is updated during the learning process. The building blocks of neural networks, perceptrons (Rosenblatt, 1958), were first modeled after the visual system for the purpose of image recognition, however they have evolved into being utilized for solving all kinds of tasks. As the use of neural networks progressed, networks became more complex with many hidden layers between the input and output of the network. These networks were referred to as “deep” learning networks due to the depth of the network’s layers and were much more powerful than their “shallow” counterparts (Dechter, 1986). With the growth of deep learning approaches to solving problems, naturally it became implemented in the medical field for problems such as segmentation in medical imaging. The use of deep learning in medical imaging aids in both reducing the subjectivity of decisions by different experts as well as reducing the amount of time required for experts to spend on each case. These benefits have the potential to provide major improvements in diagnosis, treatment planning, and follow-up of individual patients (Menze et al., 2015).

Many previous studies have incorporated deep learning in segmentation of brain tumors, especially with the introduction of several Brain Tumor Segmentation (BraTS) challenges over the past decade put forth jointly by the Radiological Society of North America (RSNA), American Society of Neuroradiology (ASNR), and Medical Image Computing and Computer Assisted Interventions Society (MICCAI), together referred to as RSNA-ASNR-MICCAI. With the popularity of the challenge, numerous different approaches to deep learning-based segmentation have been proposed with various network architectures including more simplistic convolutional neural networks (Pereira et al., 2016), fully connected conditional random fields (Kamnitsas et al., 2017), U-nets (Yang et al., 2020), and encoder-decoder networks (Rehman et al., 2021).

While deep learning in medical image segmentation can be very useful, it also faces the issue of requiring large amounts of manually annotated data to serve as the ground truth reference during training. With segmentations for some tasks becoming more complex and requiring higher levels of accuracy with fewer errors, this ground truth need grows even further. Manually annotating ground truth data can be a huge burden, requiring the valuable labor effort and cost of trained experts. Manual segmentation of brain tumors such as high-grade gliomas, for example, can take roughly 16 minutes per scan (Odland et al., 2015) and so in a dataset of around 1000 cases the amount of time required just for preparing the manually segmented dataset can take hundreds of hours. For some complex tasks, manually annotating enough data for training can become unfeasibly burdensome. For this reason, approaches to reduce the burden of acquiring adequate ground truth data for training deep learning algorithms is highly desirable.

1.5 Active Learning

A potential approach toward reducing the ground truth burden is through the implementation of active learning techniques. While often in machine learning the learner is a passive recipient of data to be processed, this "passive" role neglects the possibility of using feedback from the model to the learner's benefit in an "active" role (Cohn et al., 1996). Active learning utilizes a semi-supervised approach in which the learner makes queries to influence which data is selected by the oracle for updating the model. When data is queried properly, it can drastically reduce the data requirements for some learning problems and greatly improve efficiency (Angluin, 1988; Baum, 1991). In practice, active learning approaches are most beneficial in tasks that require very large datasets, often due to complexity, and have high cost and labor demands. The approach involves decreasing the amount of ground truth data needed for model training by first training the model with a smaller dataset, and then querying the unlabeled data for cases which would most benefit the model and adding them to the dataset. This approach makes use of the model's prior knowledge to determine which instances would be most informative rather than random selection of instances to add. An example schematic of this process can be seen in Figure 1.3.

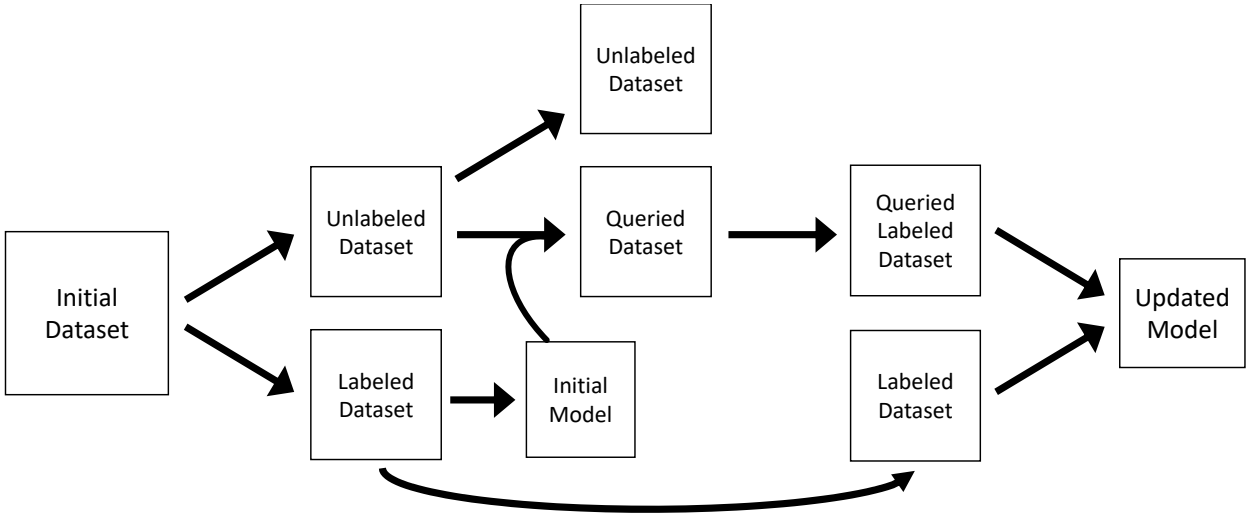

Figure 1.3. An example schematic demonstrating the active learning process.

Active learning techniques have been applied to many medical-related challenges, including classification of sleep stages (Grimova & Macas, 2019) or detecting seizures (Ge et al., 2021) from electroencephalogram (EEG), surgical workflow analysis (Bodenstedt et al., 2019), classifying cancer pathology reports (De Angeli et al., 2021), generating synthetic computed tomography (CT) images from MR data (Qian et al., 2020), and whole brain segmentation (Sourati et al., 2018). While the implementation of active learning techniques for deep learning in medicine is growing, the application of active learning in medical imaging and especially in deep learning-based segmentation is very sparse, with only a handful of studies. This study aimed to assess the application of an active learning approach to development of a deep learning-based brain glioma segmentation model from MR images.

Chapter 2

Materials and Methods

2.1 BraTS Dataset

The dataset used in this study was the publicly available training dataset provided for the 2021 RSNA-ASNR-MICCAI BraTS Challenge (Baid et al., 2021; Bakas et al., 2017; Menze et al., 2015). This dataset consisted of 1251 multi-institutional, multi-parametric MRI scans with pathologically confirmed glioma diagnosis. Each individual case contained pre-contrast T1-weighted MRI (T1), post-Gd contrast T1-weighted MRI (T1c), non-contrast T2-weighted MRI (T2), and non-contrast T2 Fluid Attenuated Inversion Recovery MRI (FLAIR), as well as a ground truth manual segmentation. An example of all images provided for a single case can be seen in Figure 2.1. The manual segmentation of ground truth data was performed after pre-processing steps including co-registration to the same SRI24 anatomical template (Rohlfing et al., 2010), resampling to a resolution of $1mm^3$, and skull-stripping. Manual segmentations were created by one of a group of neuroradiological experts and were subsequently reviewed for consistency and compliance with the annotation protocol by a single board-certified neuroradiologist with more than 15 years of experience (Bakas et al., 2018). The segmentations were done using region growing techniques interpolating between every third axial slice and were done using 3D Slicer software (Fedorov et al., 2012) with a per-subject time of approximately 60 minutes (Menze et al., 2015).

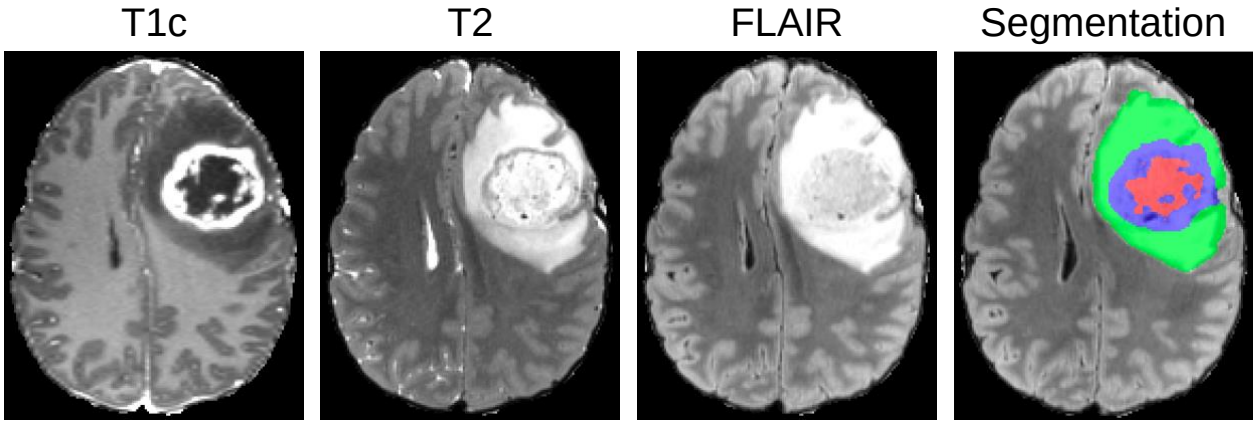

Figure 2.1. Representative images of from left to right: T1c, T2, FLAIR, and manual ground truth segmentation (red: NCR, blue: ET, green: ED) for a representative case of the BraTS 2021 dataset.

Gliomas were divided into three image-based sub-regions for segmentation: Gd-enhancing tumor, necrotic core, and peritumoral edematous/invaded tissue. Segmentation of the Gd-enhancing tumor (ET) was performed by thresholding T1c intensities within the gross tumor core on a case-by-case basis and included the Gd-enhancing tumor rim while excluding

the necrotic center and vessels (Menze et al., 2015). The segmentation of the necrotic core (NCR) was defined as tortuous, low intensity necrotic structures within the enhancing rim and was visible on T1c (Menze et al., 2015). Additionally, since 2017 the NCR label has been combined with the non-enhancing core of the tumor that remained of the gross tumor core after the ET and NCR were segmented (Bakas et al., 2018). Segmentation of the peritumoral edematous/invaded tissue (ED) was done primarily using the T2 images, followed by a cross-check with FLAIR images (Menze et al., 2015).

For the purposes of this study, the three glioma segmentation labels (ET, NCR, and ED) were combined into a single label delineating the whole tumor (WT). An example of this change can be seen in Figure 2.2. While having a tumor segmented into its various histological sub-regions has more clinical relevance than a whole-tumor segmentation, the focus of this study was on assessing active learning concepts rather than developing a high-performance segmentation model. Therefore, the more simplified whole-tumor segmentation allowed the focus to remain on active learning. With a reduced complexity of the segmentation model, less time and effort were needed for segmentation model training.

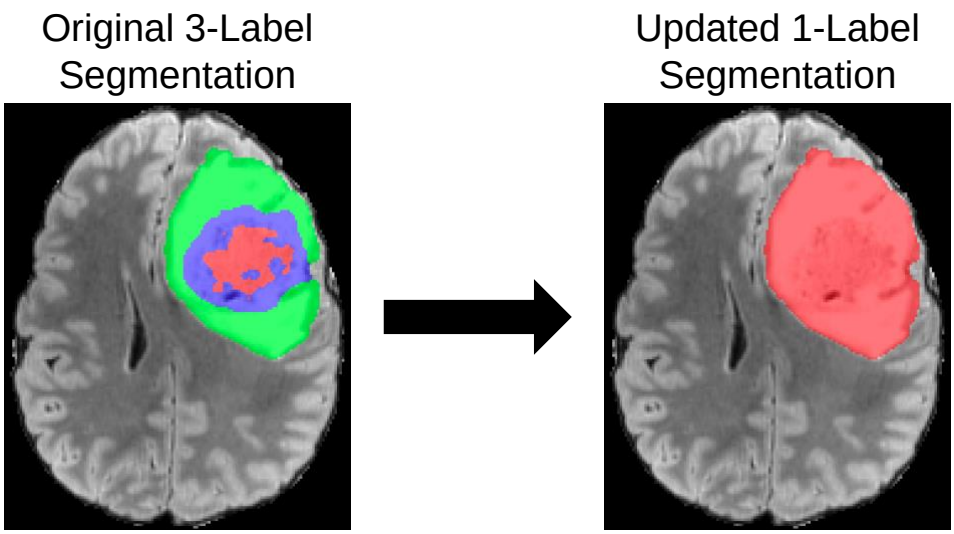

Figure 2.2. An example case demonstrating the union of the three segmentation labels for NCR (red), ET (blue), and ED (green) into a single segmentation label of WT.

2.2 Data Preprocessing

Additional preprocessing of the dataset was done upon receival to prepare the dataset for machine learning. First, the images were cropped from 240×240×155 voxels to 160×216×128 voxels to remove excess blank space in the image using a maximum intensity projection (MIP) of the full dataset to determine cropping dimensions. The images were then normalized between 0 and 1 to avoid any intensity value biases. Normalization was done using the 98th percentile value to reduce the impact of outlier voxels and prevent the intensity distributions from being skewed. Additionally, the images underwent bias field correction using the N4ITK algorithm (Tustison et al., 2010). Due to processing capacity of the NVIDIA Quadro K5000 GPU with 4GB GDDR5 RAM (NVIDIA, Santa Clara, USA), the images were then resampled

to 48×64×40 voxels using nearest neighbor interpolation. Voxel size was kept constant at 1×1×1 centimeter.

2.3 Training Reference and Baseline Models

The 1251 cases were split into a training set consisting of 1151 cases and a testing set, further referred to as T100, consisting of 100 cases. For network architecture of the baseline reference model, the state-of-the-art, high resolution, 3D convolutional HighRes3DNet (He et al., 2016; Li et al., 2017) network was used in the open-source platform NiftyNet (Gibson et al., 2018). HighRes3DNet was designed with the purpose of parcellating neuroanatomical structures from brain MRIs and is well suited to the task of this study. The network utilizes dilated convolutions and residual connections and contains 20 layers of convolutions (Li et al., 2017). The HighRes3DNet structure can be visualized in Figure 2.3. To capture low-level image features such as edges and corners, the first seven layers contain 3×3×3 voxel convolutions. The subsequent convolutional layers are dilated first by a factor of 2 then by a factor of 4 to capture mid-level and high-level image features (Li et al., 2017). Each convolutional layer is paired with a rectified linear unit (ReLU) layer and a batch normalization layer, and every two convolutional layers are grouped by residual connections. A final softmax layer provides classification scores for each voxel in the image.

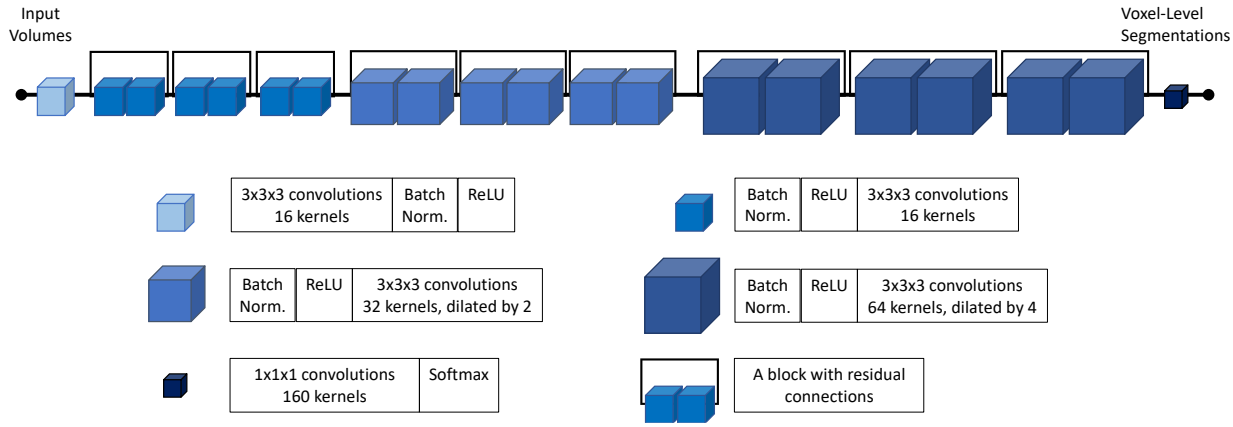

Figure 2.3. Network architecture of the HighRes3DNet used for development of the segmentation model.

A baseline model trained on all 1151 training cases (later referred to as "Model A") was trained for 64 epochs to use as a reference for comparison with models trained through active learning techniques. The training of the reference model utilized the HighRes3DNet described above. Training parameters are listed in Table 2.1. In addition to the reference model trained on all 1151 training cases, two additional baseline models were trained: one using half of the training dataset (575 cases; Model B) and the other using only 100 training cases (Model C). These two models were trained using the same protocol and parameters as the reference model. The three baseline models are summarized in Table 2.2.

Parameter	Value
Spatial Window Size The size of the input window for the image to the network	48 x 64 x 1
Optimizer A function that updates the weight parameters to minimize the loss function	Adam
Batch Size A scalar indicating the number of images to be processed in each iteration	40
Initial Learning Rate A positive scalar for updating model parameters during optimization loops	0.01
Activation Function A function that determines how the weighted sum of inputs is transferred to the output	PReLU
Loss Function A function used to compute the difference between the input and the output	Dice
Decay A scalar used to determine the strength of regularization	0.00001

Table 2.1. Training parameters for the glioma segmentation model.

Model	Training Cases	Unused Training Cases
Model A	1151	0
Model B	575	576
Model C	100	1051

Table 2.2. A summary of the number of training cases used in each of the three baseline models.

2.4 Evaluation of Baseline Models

After training of the baseline models, predicted segmentations of T100 were inferenced for each model. From these predicted segmentations and the ground truth segmentations, various evaluation metrics could be calculated to determine the agreement between the predicted and ground truth segmentations. These metrics included Sensitivity, Positive Predictive Value (PPV), Dice Similarity Coefficient (further referred to as Dice score), Jaccard Similarity Coefficient, and Modified Hausdorff Distance. For final model selection, the average Dice score of the T100 cases was computed for each iteration in each model. The "best" iteration for a given model was determined as the iteration with the maximum average Dice score and from here the rest of the metrics of this iteration were reported. Each voxel in the image can then be classified as either true positive (TP), true negative (TN), false positive (FP), or false negative (FN). The sensitivity is defined as the ratio of TP to the combined TP and FN and can be seen in equation 2.1. In other words, sensitivity describes the ability of the model to correctly

identify voxels belonging to the glioma. Sensitivity is often accompanied by specificity, however because specificity is dependent on the volume of the glioma, which varies greatly from patient to patient, it does not convey any useful information (Hatt et al., 2017). An alternative to specificity that can be used is PPV, which is defined as the ratio of TP to the combined TP and FP seen in equation 2.2. This describes the proportion of correctly identified glioma voxels.

$$Sensitivity = \frac{|TP|}{|TP| + |FN|} \tag{2.1}$$

$$PPV = \frac{|TP|}{|TP| + |FP|} \tag{2.2}$$

The Dice score is defined as twice the overlapping voxels of the segmentation and ground truth divided by the combined number of voxels of each and shown in equation 2.3. This value ranges between 0 and 1 and represents the proportion of overlap between the predicted and ground truth segmentations. A value closer to 1 suggests more overlap between the predicted and ground truth segmentations and is therefore preferred. Jaccard Similarity Coefficient is defined as the size of the intersection between the segmentation and ground truth divided by the size of the union of the two and is shown in equation 2.4. Jaccard Similarity Coefficient, like the Dice Score, ranges from 0 to 1 and explains the similarity between the two sets of segmentation voxels with values closer to 1 being preferred.

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{2.3}$$

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.4}$$

Hausdorff Distance is defined as the maximum of the minimum distances between two sets of points (*A* and *B*) in space. It is shown in equation 2.5 for which set *A* is rewritten as a_i and set *B* as b_j. The distance between voxels a_i and b_j is then denoted as $\delta(a_i, b_j)$ as the Euclidian Distance between the center of a_i and center of b_j (Hatt et al., 2017). It describes the distance between two sets of voxels and so a smaller Hausdorff Distance is preferred. Because of the nature of the equation, Hausdorff Distance becomes very sensitive to noise (Hatt et al., 2017). This issue can be addressed using a Modified Hausdorff Distance which replaces the maximum distance with average distance (Aspert et al., 2002). The equation for Modified Hausdorff Distance can be seen in equation 2.6.

$$HD(A, B) = \max\left\{max_i\ min_j\ \delta\left(a_i, b_j\right), max_j\ min_i\ \delta\left(a_i, b_j\right)\right\} \tag{2.5}$$

$$MHD(A,B) = \frac{1}{|A|}\sum_i min_j\ \delta\left(a_i, b_j\right) + \frac{1}{|B|}\sum_j min_i\ \delta\left(a_i, b_j\right) \quad (2.6)$$

2.5 Active Learning

For the implementation of active learning techniques after development of the three baseline models, Dice score was used to evaluate which cases the model performed well with and which ones the model performed poorly with. For the two models with reduced training set size, Dice score was used to determine which additional cases would be beneficial to the training of the model and would need manual segmentation by an expert. With Model B, an iteration of the model shortly after the performance began to plateau was selected for continuing with active learning. Using this model iteration, the resulting predicted segmentations were inferenced for the unused 576 training cases and the Dice scores for these predicted segmentations were computed. This earlier iteration was selected rather than the final model iteration so that any changes in model performance could be attributed to the adjustments of the training dataset rather than simply further training time. The resulting segmentations were then dichotomized based on their Dice scores into two categories: above or below a threshold score of 0.7. The threshold of 0.7 was selected from literature on image validation suggesting Dice scores above 0.700 are considered to have a good overlap (Zijdenbos et al., 1994). For those above a 0.7 Dice score, the segmentations were considered as "acceptable" and that the model did not struggle with the case and so the case was addended to the training dataset with the predicted segmentation in place of the ground truth. These cases represented those that would not require manual segmentation because the model had already learned how to adequately segment these images. For those below a 0.7 Dice score, the segmentations were considered as "poor" predictions that the model struggled with. These cases were addended to the training dataset but using the original ground truth segmentation and represented cases that required manual segmentation, as the model was still having trouble with the segmentation predictions. The process of replacing the predicted segmentation with the ground truth for these cases was equated to an expert manually segmenting/adjusting the case. Model training then continued with the updated dataset of a combined 575 initial training cases and 576 addended training cases for a total of 1151 training cases. The same procedure was followed for Model C using 200 additional training cases, constituting one round of active learning. For Model C, the process was repeated twice more with 600 additional training cases for an updated size of 900 training cases followed by 251 additional training cases to increase the size to the full dataset of 1151 training cases. The number of training cases used in each active learning model is summarized in Table 2.3.

Model	Training Cases	Unused Training Cases
Model B	575	576
Model B Active Learning	1151	0
Model C	100	1051
Model C Active Learning 1	300	851
Model C Active Learning 2	900	251
Model C Active Learning 3	1151	0

Table 2.3. A summary of the number of training cases used in each active learning model.

2.6 Secondary Classification Model

In a real-world scenario, access to ground truth segmentations is limited and so computing Dice scores on unused training data is not feasible and any data with available ground truth data would be used in the training of the segmentation model itself. To address this issue, a secondary model was developed to predict the segmentation quality when provided with the MR images and their predicted segmentation without the aid of the ground truth segmentation. Input to this model included the T1c, T2, and FLAIR images as well as the segmentation probability map output of the segmentation model after the softmax layer. Rather than voxels being integer values depending on their predicted class of glioma or not glioma, the segmentation probability map takes the output from one step back when each voxel is represented by a value between 0 and 1 representing the probability of that voxel to belong to a given class, therefore providing extra information giving insight into the confidence of the model in its predicted segmentation. For this reason, the segmentation probability map was used in place of the predicted segmentation. Example input images can be seen in Figure 2.4. The Dice scores were separated into three classes: “Poor Quality” for those below 0.6 Dice score, “Acceptable with Adjustments” for those between 0.6 and 0.8 Dice score, and “Acceptable Quality” for those above 0.8 Dice score. The 0.8 Dice score threshold was selected as the average Dice score of all models submitted to the 2021 BraTS Challenge and the 0.6 Dice score was selected as being slightly above the central 0.5 Dice score. A predicted segmentation in the “Poor Quality” class would need to be segmented completely manually by an expert, a segmentation in the “Acceptable with Adjustments” class would need some minor adjustments by an expert before being accepted, and a segmentation in the “Acceptable Quality” class could be accepted with possibly only a brief visual check by an expert. Predicted segmentations and Dice scores were taken from the iteration of Model B used for active

learning. The unused 576 training cases were used for training of the secondary classification model as to not use cases that the initial model was trained on. Additionally, due to the performance of the initial segmentation model, the Dice scores of the training set were skewed towards higher Dice scores. To counter this the amount of training cases for each class was balanced, resulting in 78 cases per class for a total of 234 training cases. For consistency, the testing set used for this classification model was the same T100 set used for testing the initial segmentation models.

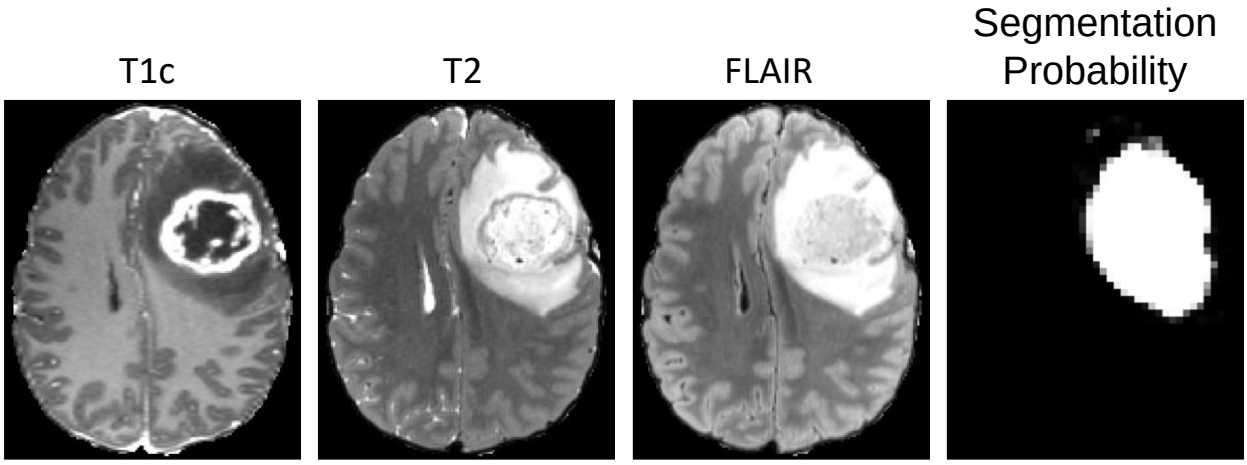

Figure 2.4. Example input images for the secondary classification model including from left to right: T1c, T2, FLAIR, and the predicted segmentation probability map.

2.7 Training Secondary Classification Model

For the secondary classification model, a model was trained in the Matlab R2022a Deep Learning Toolbox (MathWorks, Portola Valley, United States) for 60 epochs with an initial learning rate of 0.1 using a piecewise schedule that dropped the learning rate by a factor of 0.1 every 10 epochs. The network consisted of 61 total layers, beginning with a 3D input layer followed by 14 blocks consisting of a max pooling layer, convolution layer, batch normalization layer, and ReLU layer. The first of these blocks did not have a max pooling layer. For the first 7 blocks, the subsequent convolutions were dilated by a factor of 2 and for the remaining blocks they were contracted by a factor of 2. Following these blocks, the network ended with a dropout layer with a value of 0.1, two fully connected layers of size 512 and 3, respectively, and a softmax layer before the final classification layer. The network architecture can be seen in Figure 2.5. Training parameters for the classification model can be seen in Table 2.4.

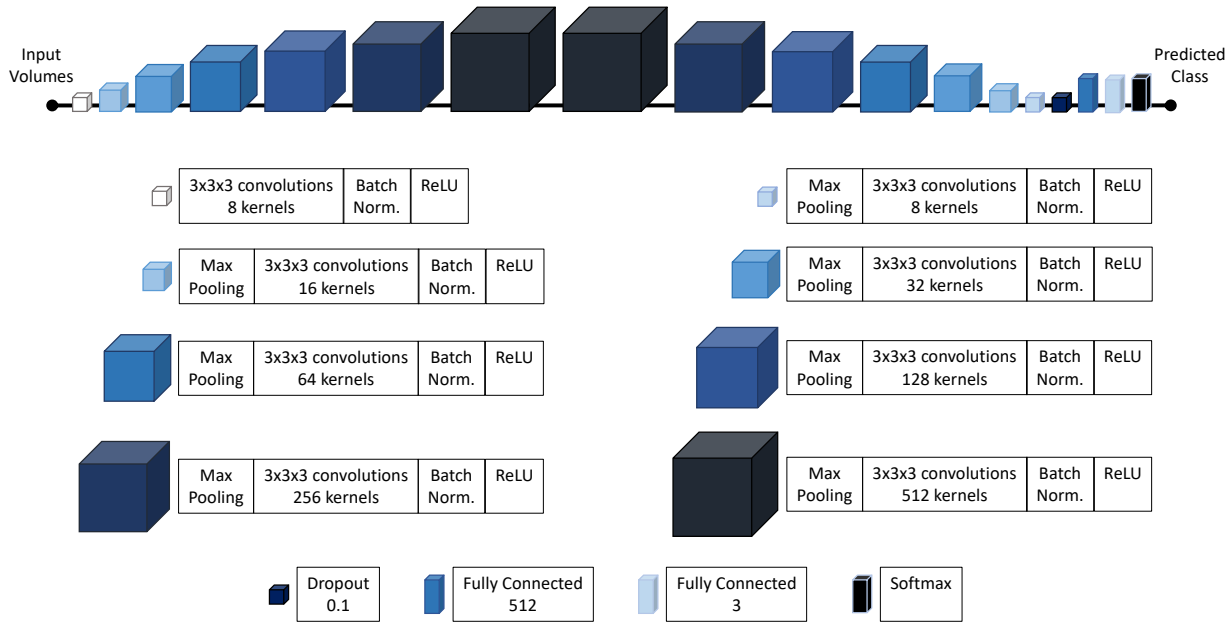

Figure 2.5. Network architecture of the secondary classification model for classifying predicted segmentation quality.

Parameter	Value
Optimizer A function that updates the weight parameters to minimize the loss function	SGDM
Batch Size A scalar indicating the number of images to be processed in each iteration	64
Initial Learning Rate A positive scalar for updating model parameters during optimization loops	0.1
Momentum A scalar from 0 to 1 of the contribution of the previous step to the current step	0.9
L2 Regularization A non-negative scalar of weight decay for the loss function to prevent overfitting	0.00005
Gradient Threshold A positive scalar that clips the gradient whenever the value is exceeded	0.5

Table 2.4. Training parameters for the classification model.

The results of the classification model were evaluated using the sensitivity, specificity, PPV, F-Score, and area under the receiver operating characteristic curve (AUC). F-score ranges from 0 to 1 and is calculated from the precision (also called PPV) and recall (also called sensitivity) as their harmonic mean (Taha & Hanbury, 2015). The equation for F-score can be seen in Equation 2.7. The AUC represents the probability of the classifier to rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2004). Despite ranging from 0 to 1, a perfectly random guessing results in a diagonal line with an AUC of 0.5 and so the more realistic range for an AUC is instead from 0.5 to 1.

$$FScore = \frac{2*PPV*Sensitivity}{PPV+Sensitivity} \quad (2.7)$$

Chapter 3

Results

3.1 Baseline Model Metric Results

Metric results for the "best performing" iteration of each of the three baseline models are summarized in Table 3.1, with the iteration selected through the Dice score due to this metric being used for thresholding further in the study. The iterations selected for each model were the 52nd epoch for Model A with an average Dice score of 0.906, the 56th epoch for Model B with an average Dice score of 0.865, and the 57th epoch for Model C with an average Dice score of 0.825. For sensitivity, Models A and B had similar values of 0.912 and 0.913, respectively, while Model C had the lowest with 0.849. For the rest of the metrics, Model A showed the best values while Model C showed the worst, and with Model B somewhere in the middle between the two. For PPV, values were 0.906, 0.842, and 0.825 for Models A, B, and C, respectively. Dice scores were 0.906, 0.865, and 0.825 for Models A, B, and C, respectively, and Jaccard Similarity Coefficients were 0.834, 0.778, and 0.718 for Models A, B, and C, respectively. Model A showed the shortest Modified Hausdorff Distance of 3.309, followed by Model B with 3.710, and finally Model C with 4.317.

Model	Epoch	Sensitivity	Positive Predictive Value	Dice Similarity Coefficient	Jaccard Similarity Coefficient	Modified Hausdorff Distance
Model A; 1151 Training Cases	52	0.912	0.906	0.906	0.834	3.309
Model B; 575 Training Cases	56	0.913	0.842	0.865	0.778	3.710
Model C; 100 Training Cases	57	0.849	0.825	0.825	0.718	4.317

Table 3.1. Metric results for each baseline model.

The iterations selected for each model for active learning purposes were epoch 23 for Model B with a Dice score of 0.854 and epoch 23 for Model C with a Dice score of 0.813. The iterations selected for the following rounds of active learning for Model C were epoch 38 with a Dice score of 0.830 for the 2nd round and epoch 50 with a Dice score of 0.841 for the 3rd round. Full metrics for these iterations can be seen in Table 3.2. Of the unseen dataset in Model B, 127 of the 576 cases (22.0%) were below the threshold of 0.7 Dice score and were replaced

with the ground truth image. For the first round of active learning with Model C, 43 of the 200 unseen cases (21.5%) were below the 0.7 Dice score threshold. In the following rounds of active learning for Model C, 133 of the 600 unseen cases (22.2%) and 53 of the 251 unseen cases (21.1%) were below the threshold and replaced by ground truth segmentations for the 2nd and 3rd rounds of active learning, respectively.

Model	Epoch	Sensitivity	Positive Predictive Value	Dice Similarity Coefficient	Jaccard Similarity Coefficient	Modified Hausdorff Distance
Model B; Active Learning Iteration	23	0.872	0.859	0.854	0.757	4.045
Model C; Active Learning Iteration 1	23	0.851	0.803	0.813	0.702	4.413
Model C; Active Learning Iteration 2	38	0.869	0.815	0.830	0.725	4.215
Model C; Active Learning Iteration 3	50	0.876	0.829	0.841	0.741	4.067

Table 3.2. Metric results for each model iteration selected for the active learning process.

3.2 Training Process

The following figures display the training progress of each model's Dice score. In Figure 3.1, the progress of the three baseline models is displayed. Model A (blue) shows the highest Dice score consistently through its training. Model B (orange) takes the longest for its Dice score to begin to plateau, however the plateaued Dice scores are still above those of Model C (grey). The active learning progress of Model B is displayed in Figure 3.2, in which the baseline (orange) and active learning (yellow) models are shown. The active learning model shows higher peak Dice score than that of the baseline model. The same is shown for Model C in Figure 3.3. With each round of active learning for Model C, the Dice scores show improvement. The baseline (grey) has the lowest peak Dice score, followed by the first round of active learning (light blue), the second round of active learning (green), and finally the third round of active learning (navy) with the highest Dice scores.

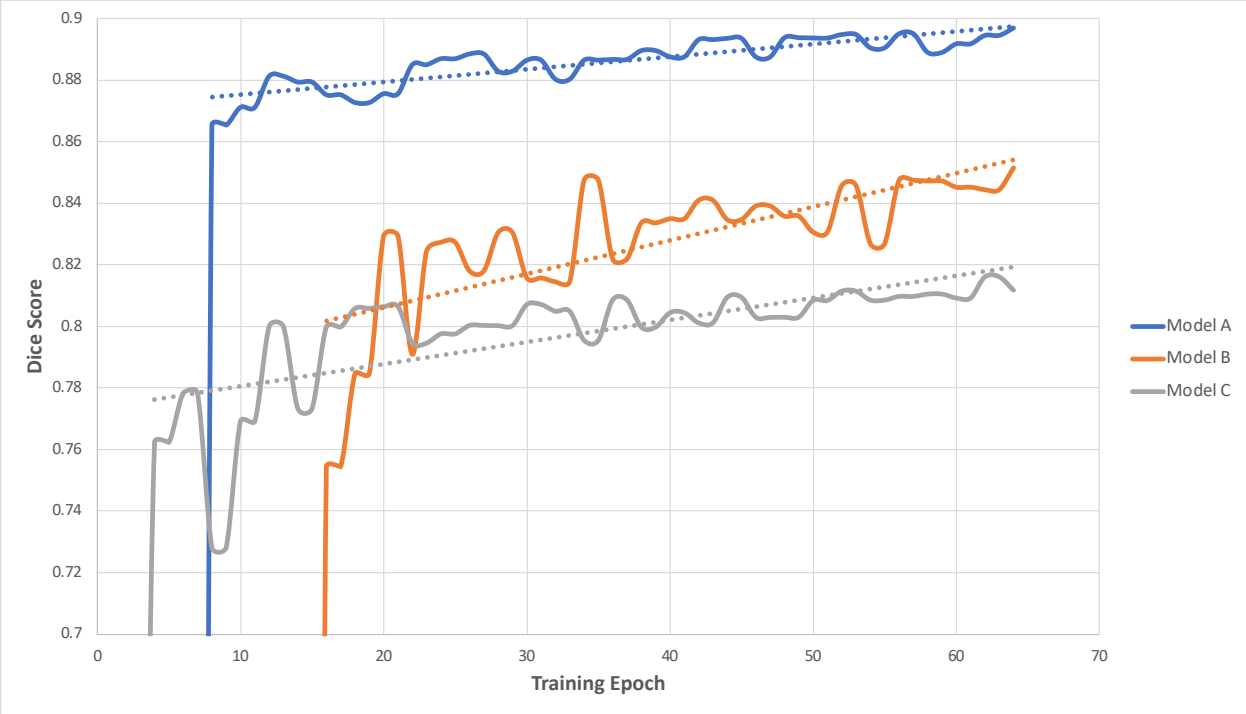

Figure 3.1. Training progress of the three baseline models.

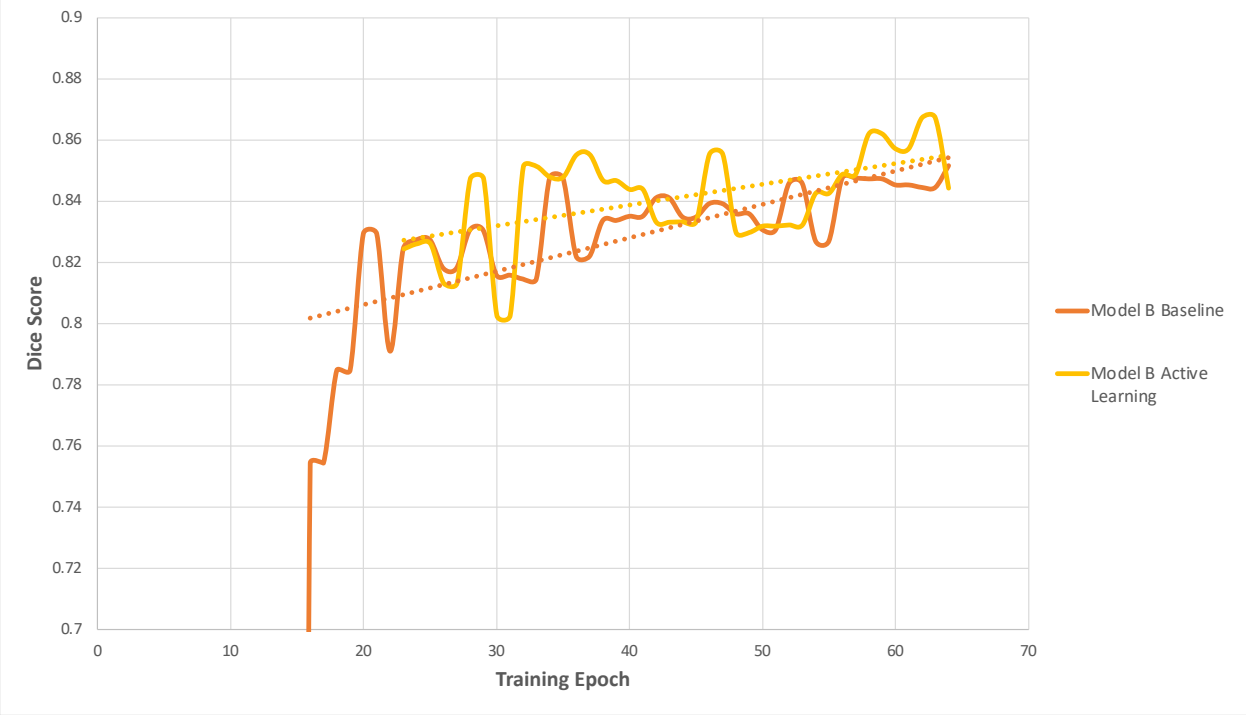

Figure 3.2. Training progress of Model B's baseline and active learning models.

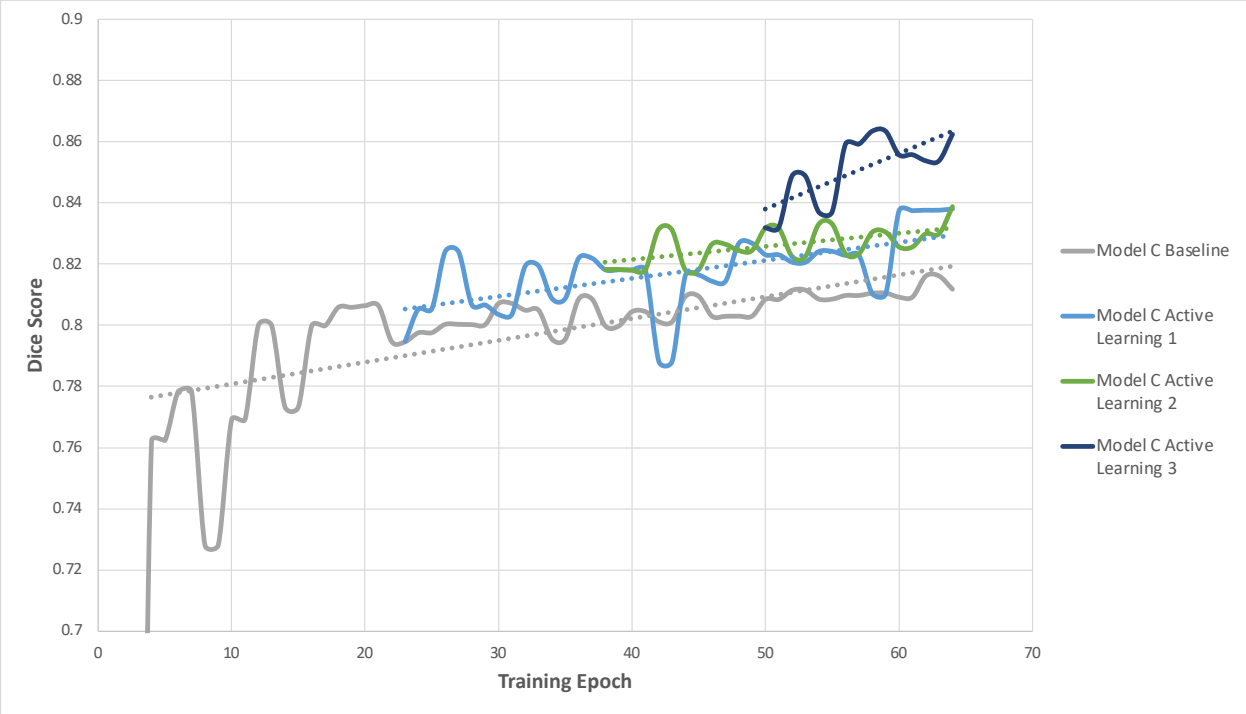

Figure 3.3. Training progress of Model C's baseline and active learning models.

3.3 Active Learning Results

The final resulting metrics of the segmentation models can be seen in Table 3.3, including the reference Model A, the baselines of Models B and C, and the final post-active learning results for Models B and C. In terms of peak average Dice scores, the post-active learning models of Models B and C both showed improvement, with the overlap between the predicted and ground truth segmentations increasing an average of 0.5% for Model B and 4.3% for Model C. Though the improvements for Model B were not as pronounced, Model C's improvements from pre- to post-active learning were more notable. Additionally, through active learning Model C's sensitivity improved from 0.849 to 0.907 and PPV from 0.825 to 0.845. The intersection between the predicted and ground truth segmentations represented by the Jaccard Similarity Coefficient also increased through active learning from 0.718 to 0.777, and the Modified Hausdorff Distance showing the distance between the two sets of voxels from the predicted and ground truth segmentations improved with a decrease of 12.9%.

Model	Epoch	Sensitivity	Positive Predictive Value	Dice Similarity Coefficient	Jaccard Similarity Coefficient	Modified Hausdorff Distance
Model A; Reference	52	0.912	0.906	0.906	0.834	3.309
Model B; Baseline	56	0.913	0.842	0.865	0.778	3.710
Model B; Post-Active Learning	63	0.882	0.876	0.870	0.780	3.837
Model C; Baseline	57	0.849	0.825	0.825	0.718	4.317
Model C; Post-Active Learning	57	0.907	0.845	0.868	0.777	3.761

Table 3.3. Metric results of the pre-active learning and post-active learning segmentation models.

For qualitative visual assessment of segmentation prediction, a representative example case of segmentations before and after active learning from Model C can be seen in Figure 3.4 for the three MR sequences. In the example case, before active learning the model predicted a large region opposite the glioma to be glioma tissue. Visually, the incorrectly segmented region appears larger than the ground truth segmentation. After active learning, this incorrectly segmented region has disappeared and only the true glioma has been segmented. For additional qualitative assessment in Figure 3.5, two example cases from Model C after the active learning process are displayed representing low Dice score and high Dice score cases. In the high Dice score case with an individual case Dice score of 0.96, the glioma is very prominent in the MR images both in terms of size and contrast with the background brain tissue. In the low Dice score case with an individual case Dice score of 0.62 on the other hand, the glioma is much smaller and less prominent in the MR images. The predicted segmentation by the model over-segments the glioma, capturing healthy brain tissue in the segmentation.

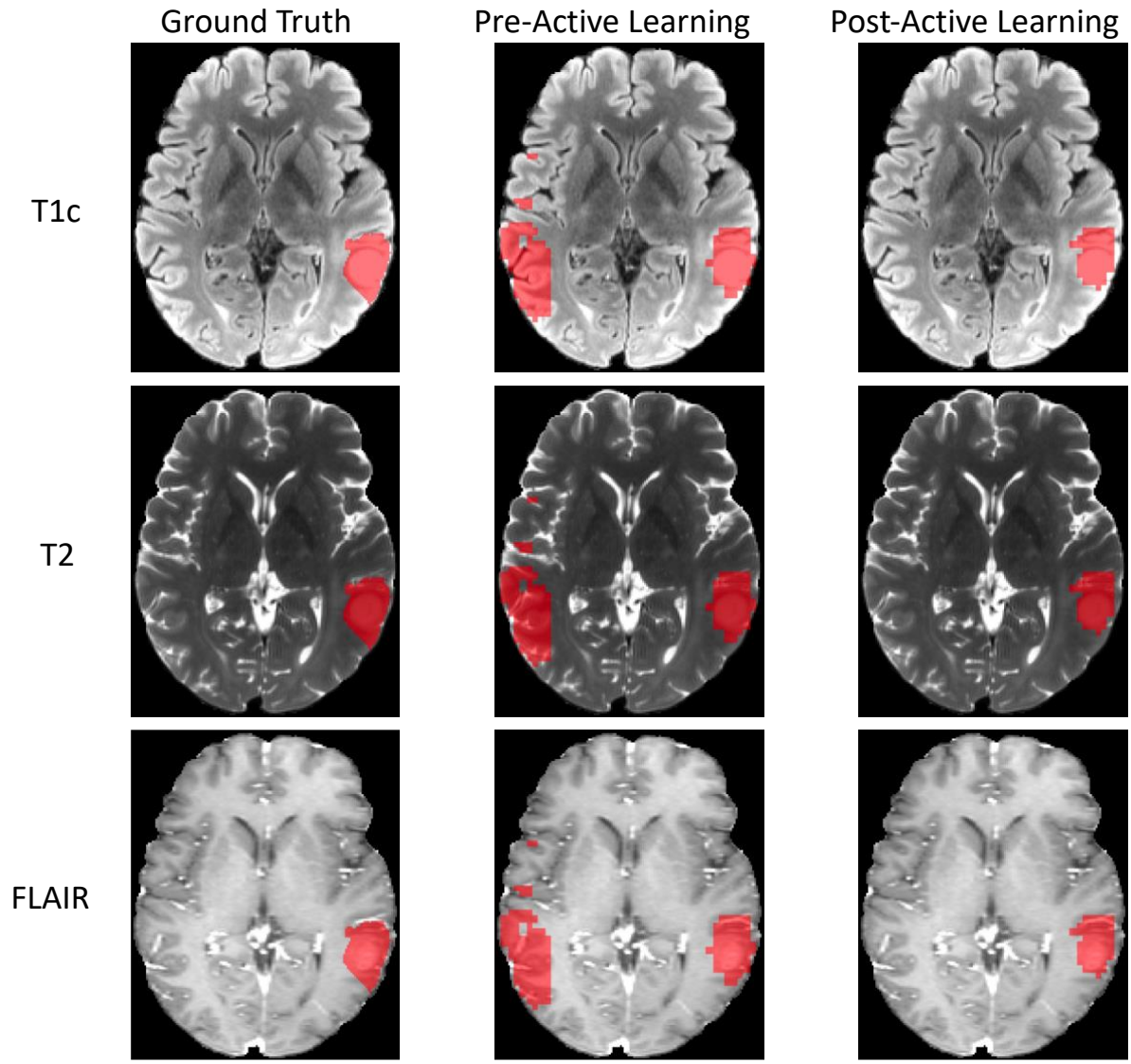

Figure 3.4. Ground truth, pre-, and post-active learning predicted segmentations for T1c, T2, and FLAIR images for a representative case in Model C.

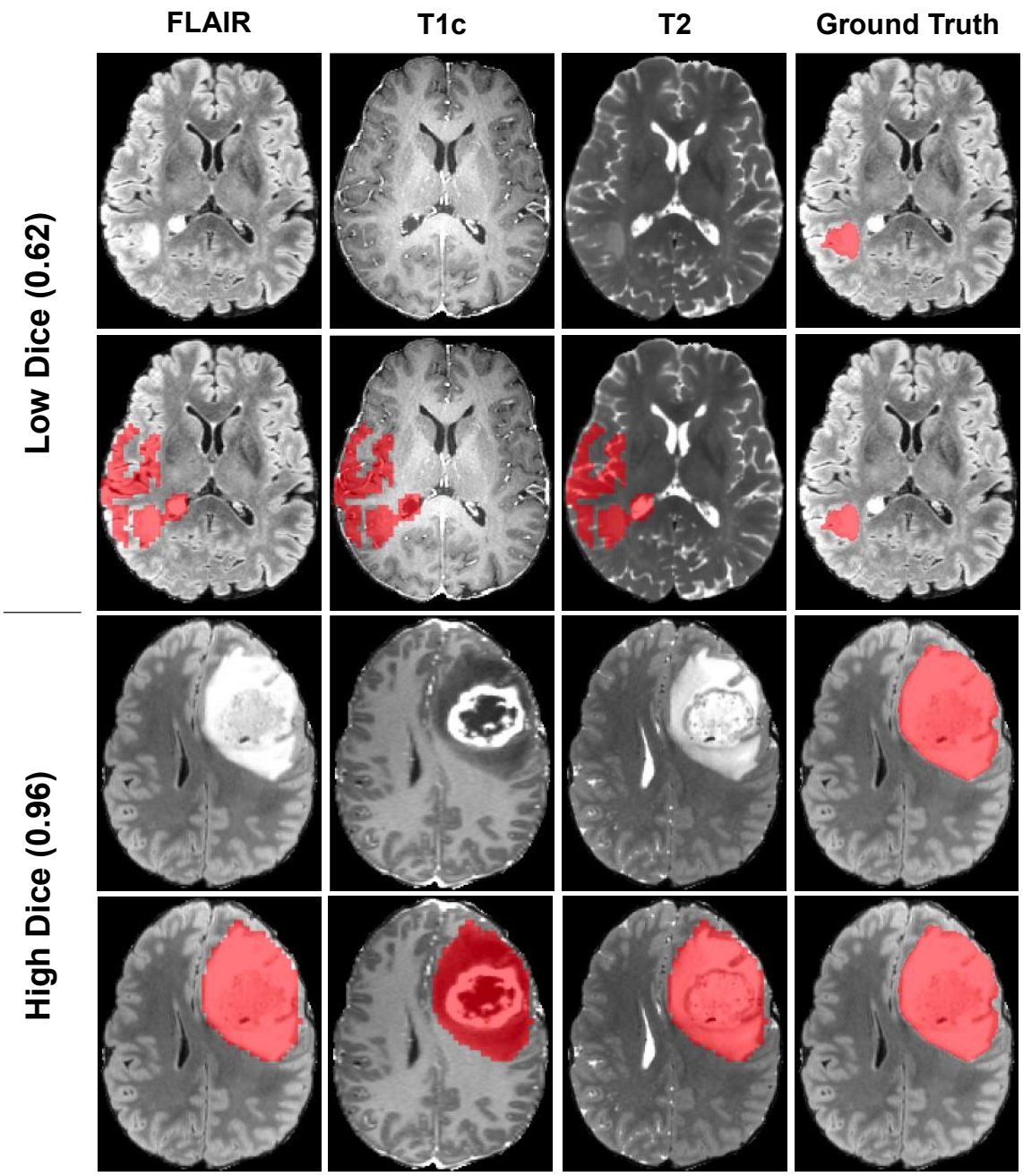

Figure 3.5. Example cases from Model C with low (above) and high (below) Dice scores and their predicted and ground truth segmentations for T1c, T2, and FLAIR.

3.4 Classification Model Results

A confusion matrix of the model to classify predicted segmentations can be seen in Figure 3.6 for the classes "Poor Quality" in which a physician needs to segment the image from scratch, "Acceptable with Adjustments" in which a physician needs to check and edit the segmentation, and "Acceptable Quality" in which the segmentation does not require any checking or editing by a physician. Of 100 total T100 cases, 82 were classified to the correct class while 18 were misclassified for an accuracy of 82%. Of the four cases in the "Poor Quality" class, three (75%) were classified correctly with the incorrectly classified case assigned to the "Acceptable with

Adjustments" class. For the 11 cases in the "Acceptable with Adjustments" class, 6 (54.5%) were classified correctly while 5 were incorrectly classified to the "Acceptable Quality" class. Of the 85 cases in the "Acceptable Quality" class, 73 (85.9%) were classified correctly while 12 were incorrectly classified to the "Acceptable with Adjustments" class. Of the 18 misclassified cases, 13 (72.2%) were misclassified into a class that would still require assessment by an expert ("Poor Quality" or "Acceptable with Adjustments"). Additional metrics of the classification model results can be seen in Table 3.4. Figure 3.7 displays the ROC curves for the three predicted Dice score classes with AUC scores listed. The "Poor Quality" class of Dice scores less than 0.6 showed the highest AUC of 0.995, followed by the "Acceptable Quality" class of Dice scores above 0.8 with an AUC of 0.877, and finally the "Acceptable with Adjustments" class of Dice scores between 0.6 and 0.8 with an AUC of 0.810.

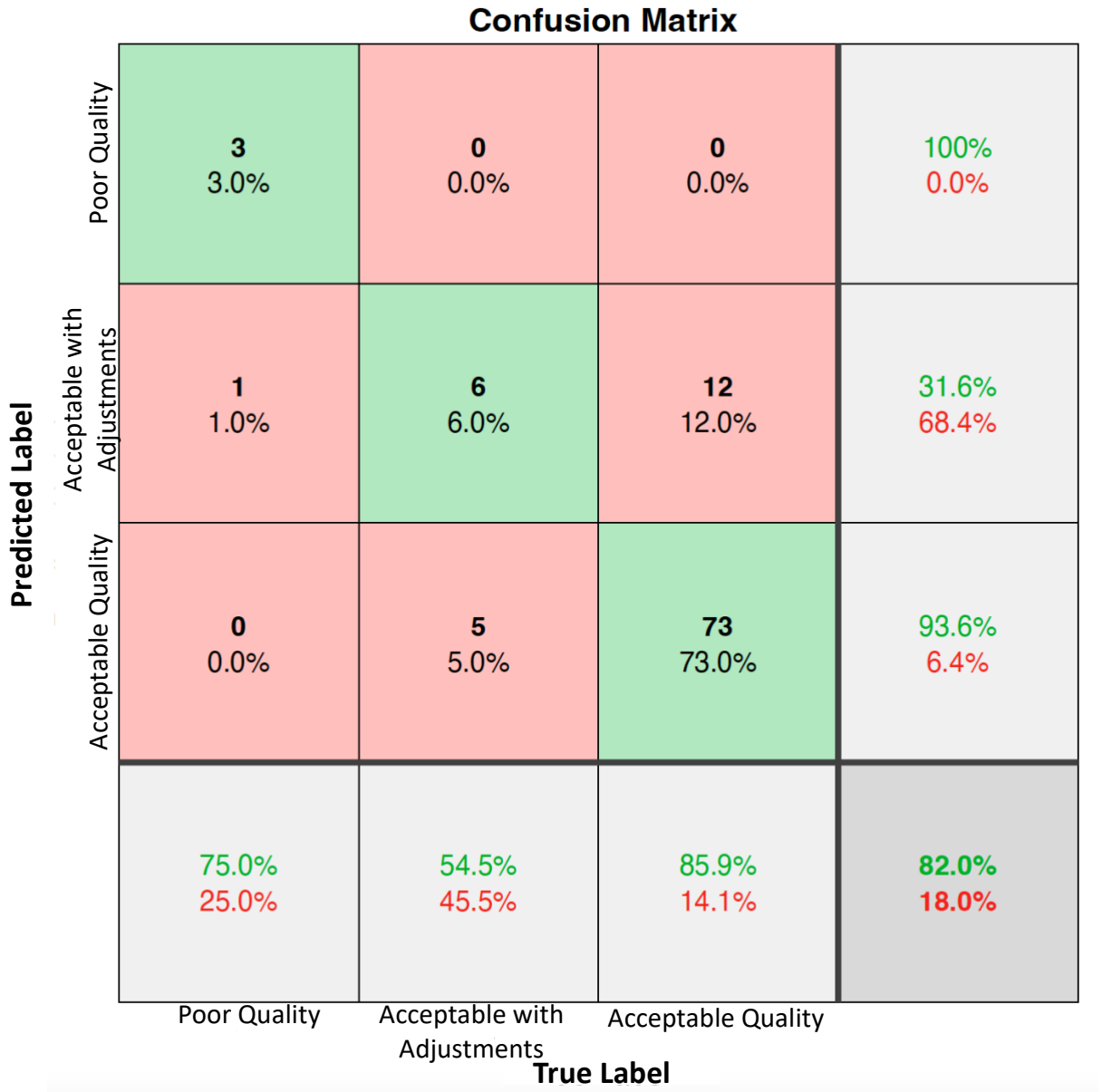

Figure 3.6. Confusion matrix for the classification of predicted segmentations into "Poor Quality", "Acceptable with Adjustments", and "Acceptable Quality".

Class	Sensitivity	Specificity	Positive Predictive Value	F-Score	AUC
Poor Quality	0.750	1.000	1.000	0.857	0.995
Acceptable with Adjustments	0.545	0.938	0.316	0.400	0.810
Acceptable Quality	0.859	0.455	0.936	0.896	0.877

Table 3.4. Metric results for each class of the classification model.

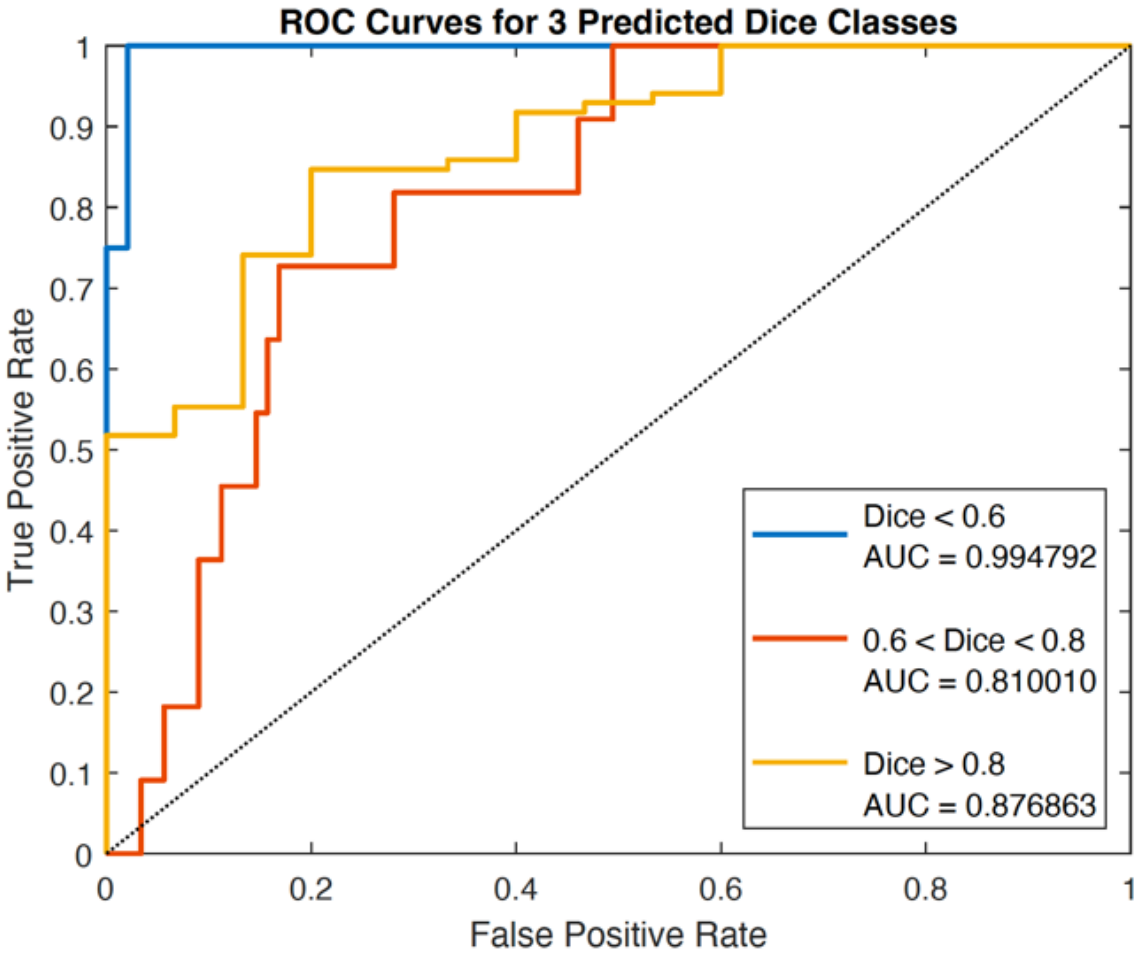

Figure 3.7. ROC curves for the three classes predicted by the secondary classification model for predicting the quality of segmentation.

Chapter 4

Discussion

In this study, the application of an active learning approach to segment whole brain gliomas from MRI was assessed. The key benefit to the active learning concept lies in its potential reduction of data requirements, with preferential data being selected for model training through feedback from the model. After three baseline segmentation models were trained as reference, active learning was applied to the two models of reduced dataset size using a Dice score threshold and the training sets were updated based on the queried data. While this first step allowed for the assessment of the viability of active learning in training glioma segmentation models as a concept, it relied on prior knowledge of the ground truth data for the unseen cases to compute Dice scores. In a clinical or real-world setting, this would not be practical as one would want to utilize all available training data that is accompanied by a ground truth segmentation to train the best model possible. A secondary classification model was then developed to address this challenge with the goal of classifying predicted segmentations into those of "Poor Quality", "Acceptable with Adjustments", and "Acceptable Quality" using Dice score thresholds of below 0.6, between 0.6 and 0.8, and above 0.8, respectively.

Results of the quantitative analysis of the segmentation models demonstrated that an active learning approach when applied to glioma segmentation from MR images shows comparable segmentation results to reference non-active learning models but at a lower ground truth cost. With active learning, the average Dice score of the predicted segmentations of T100 rose from 0.865 to 0.870 for Model B and from 0.825 to 0.868 for Model C. While these two models did not quite reach the Dice score of the reference Model A (0.906), the Dice scores were still comparably high and with much less manual segmentation required for training. For Model B, only 127 of the additional 576 cases required manual segmentation for a total of 702 of the 1151 cases. This reduced the total number of cases needing an expert's manual segmentation by 449 or 39.0% of the total training dataset. For Model C, across all 3 rounds of active learning only 229 of the additional 1051 cases required manual segmentation, reducing the number of total training cases with expert manual segmentation by 822 and meaning that only 329 or 28.6% of the 1151 cases required manual segmentation. These drastic reductions in manual segmentation required would greatly save in the cost of time and labor by trained experts. Though the segmentations through active learning did not quite reach the levels of the reference model, there is a trade-off in which the reductions of manually segmented ground truth data required can make up for this. This may be especially useful in tasks for which there is more leniency in the precision and so the slight decrease in accuracy of the predicted segmentations is less important compared with the time and effort saved.

In the second step of the study, the classification model was highly sensitive to cases with "Poor Quality" predicted segmentations of low Dice score (<0.6) which can be very useful in identifying the problematic cases. Additionally, though there were 18% of cases misclassified, many of these would not negatively impact the results. For example, though there were 12 cases of "Acceptable Quality" misclassified to be in the class for "Acceptable

with Adjustments", this incorrect class would simply suggest that an expert would need to make minor adjustments and in doing so the expert would see that the quality of the segmentation is acceptable. Therefore, being misclassified into a lower class does not negatively impact the results. There were only 6 of the 100 T100 cases misclassified into a class of better quality, with one of these being misclassified from "Poor Quality" to "Acceptable with Adjustments" which would still warrant an expert to visually assess the segmentation. Additionally, none of the misclassifications jumped between "Poor Quality" and "Acceptable Quality" and only into "Acceptable with Adjustments". When taking these instances into account, 95% of the cases were either classified correctly or if incorrectly then into a class that would still require an expert to visually assess the predicted segmentation and adjust if necessary.

Though not for glioma segmentation specifically, various other studies have also implemented active learning techniques to medical image segmentation toward reducing manual segmentation data requirements. In a study applying active learning to interactive 3D image segmentation (Top et al., 2011), an active learning technique involving uncertainty fields based on boundary, regional, smoothness and entropy terms were applied to various tasks including segmentation of the putamen from brain MRI, liver in abdominal CT, and a collection of pelvic bones and muscles in both CT and MRI. The study found that in addition to either comparable or improved Dice scores, the active learning techniques also reduced user input by an average of 64%. This finding shows a similar reduction in human effort of segmentation as the present study with a 61% reduction in Model B and a 72.4% reduction in Model C. Another study focusing on generation of realistic chest x-ray images using a conditional generative adversarial network followed by a Bayesian neural network to calculate informativeness for active learning (Mahapatra et al., 2018) similarly found that an active learning framework was able to achieve comparable results using only 35% of the full dataset. In a study of hippocampal segmentation from MR images (Nath et al., 2021), a Query-by-Committee approach to active learning was implemented and was able to achieve full segmentation accuracy using just 23% of the dataset. While these studies all show drastic reductions in data requirements, they each also use different approaches to the application of active learning concepts. This suggests two things— first, there are many different techniques to approach active learning while achieving similarly small data requirement results; second, with multiple possible techniques there may be an approach that works best for a given task and so future studies wishing to optimize the process may need to test multiple approaches.

Overall, the active learning approach in this study demonstrated substantial reductions in the required amount of manually segmented ground truth data for model training. Despite these results, however, the approach faces several limitations. First, this study primarily focused on the Dice Similarity Coefficient for assessment of the segmentation quality. This led to a one-dimensional decision on the segmentation quality rather than incorporating multiple metrics into the decision such as sensitivity, positive predictive value, Jaccard Similarity Coefficient, and Modified Hausdorff Distance. Second, due to the nature of metrics like Dice Similarity Coefficient in which it compares the predicted segmentation with the ground truth segmentation, a secondary model was required for a "real-world" scenario in which the ground truth data is not already present. This second step introduces further potential for error and so having the process condensed to a single step would be ideal. Lastly, the images were

resampled to a smaller resolution and the initial 3-label segmentation was combined to a single-label segmentation to reduce complexity. While this greatly affected the quality and useability of the final predicted segmentations from the model, the overall accuracy and strength of the segmentation result was not important in this study but rather the viability and benefit of active learning concepts in model development. Despite these limitations, this study demonstrated that active learning can greatly reduce the efforts of preparing ground truth data for training segmentation models. With the querying technique being a crucial aspect in the success of the active learning model (Ge et al., 2021; Madhawa & Murata, 2020), future studies can explore alternate techniques and potentially improve results even further.

Chapter 5

Conclusion

In this study, active learning concepts were applied to a deep learning segmentation of brain gliomas from MR images to assess their viability in reducing the required amount of manually annotated ground truth data in model training. Three models were trained, each with different training set sizes (1151, 575, and 100 cases, respectively), and segmentations of the unseen training data were predicted. The predicted segmentations with a Dice score above 0.7 were used in place of ground truth and the training sets were updated for further training. It was demonstrated that using this active learning approach, more than 60% of the dataset did not require manual segmentation for adequate training of the model, suggesting that active learning when applied to model training can drastically reduce the time and labor spent on preparation of ground truth training data. While this result showed excellent promise in the usefulness of active learning in medical image segmentation as a concept, the approach relied on the availability of ground truth segmentations to determine predicted segmentation quality and so the approach lacked real-world feasibility. To address this issue, a secondary model was developed to classify the predicted segmentations into three classes based on their quality: "Poor Quality", "Acceptable with Adjustments" and "Acceptable Quality". The model performed well and classified 82% of the cases correctly and with all AUC's above 0.8 (0.995, 0.810, and 0.877). Additionally, only 5% of the cases were misclassified into a class that would not require intervention by an expert. The results of the classifier suggested that in addition to active learning concepts being greatly beneficial toward streamlining model training for medical image segmentation tasks, approaches that do not require prior knowledge of the unseen data are feasible as well. While the accuracy of the segmentation model used in this task does not yet meet the standards for clinical use, this study serves as a baseline for future work to further delve into the application of active learning for medical image segmentation, perhaps testing other active learning techniques as well.

References

Angluin, D. (1988). Queries and Concept Learning. *Machine Learning*, *2*, 319-342.

Aspert, N., Santa-Cruz, D., & Ebrahimi, T. (2002). MESH: measuring errors between surfaces using the Hausdorff distance. *Proceedings. IEEE International Conference on Multimedia and Expo*, *1*, 705-708. https://doi.org/10.1109/ICME.2002.1035879

Baid, U., Ghodasara, S., Bilello, M., Mohan, S., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F., Pati, S., Prevedello, L., Rudie, J., Sako, C., Shinohara, R., Bergquist, T., Chai, R., Eddy, J., Elliott, J., Reade, W., & Bakas, S. (2021). The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. *arXiv*, arXiv:2107.02314.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., & Davatzikos, C. (2017). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*, *4*, 170117. https://doi.org/10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R., Berger, C., Ha, S., Rozycki, M., Prastawa, M., Alberts, E., Lipkova, J., Freymann, J. B., Kirby, J., Bilello, M., Fathallah-Shaykh, H., Wiest, R., Kirschke, J., & Menze, B. H. (2018). Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *arXiv*, arXiv:1811.02629.

Bakshi, R., Ariyaratana, S., Benedict, R. H., & Jacobs, L. (2001). Fluid-attenuated inversion recovery magnetic resonance imaging detects cortical and juxtacortical multiple sclerosis lesions. *Arch Neurol*, *58*(5), 742-748. https://doi.org/10.1001/archneur.58.5.742

Balafar, M., Ramli, A., Saripan, M., & Mashohor, S. (2010). Review of brain MRI image segmentation methods. *Artif. Intell. Rev.*, *33*(3), 261-274.

Bauer, S., Wiest, R., Nolte, L. P., & Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumor studies. *Phys Med Biol*, *58*(13), R97-129. https://doi.org/10.1088/0031-9155/58/13/R97

Baum, E. B. (1991). Neural net algorithms that learn in polynomial time from examples and queries. *IEEE Trans Neural Netw*, *2*(1), 5-19. https://doi.org/10.1109/72.80287

Bodenstedt, S., Rivoir, D., Jenke, A., Wagner, M., Breucha, M., Muller-Stich, B., Mees, S. T., Weitz, J., & Speidel, S. (2019). Active learning using deep Bayesian networks for surgical workflow analysis. *Int J Comput Assist Radiol Surg*, *14*(6), 1079-1087. https://doi.org/10.1007/s11548-019-01963-9

Chen, R., Smith-Cohn, M., Cohen, A. L., & Colman, H. (2017). Glioma Subclassifications and Their Clinical Significance. *Neurotherapeutics*, *14*(2), 284-297. https://doi.org/10.1007/s13311-017-0519-x

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active Learning with Statistical Models. *Journal of Artificial Intelligence Research*, *4*, 129-145.

Croswell, J. M., Ransohoff, D. F., & Kramer, B. S. (2010). Principles of cancer screening: lessons from history and study design issues. *Semin Oncol*, *37*(3), 202-215. https://doi.org/10.1053/j.seminoncol.2010.05.006

De Angeli, K., Gao, S., Alawad, M., Yoon, H. J., Schaefferkoetter, N., Wu, X. C., Durbin, E. B., Doherty, J., Stroup, A., Coyle, L., Penberthy, L., & Tourassi, G. (2021). Deep active learning for classifying cancer pathology reports. *BMC Bioinformatics*, *22*(1), 113. https://doi.org/10.1186/s12859-021-04047-1

De Coene, B., Hajnal, J. V., Gatehouse, P., Longmore, D. B., White, S. J., Oatridge, A., Pennock, J. M., Young, I. R., & Bydder, G. M. (1992). MR of the brain using fluid-attenuated inversion recovery (FLAIR) pulse sequences. *AJNR Am J Neuroradiol*, *13*(6), 1555-1564. https://www.ncbi.nlm.nih.gov/pubmed/1332459

Dechter, R. (1986). Learning While Searching in Constraint-Satisfaction-Problems. *AAAI-86 Proceedings*, 178-183.

Fawcett, T. (2004). ROC Graphs: Notes and Practical Considerations for Researchers. *Machine Learning*, *31*, 1-38.

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J. C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J. V., Pieper, S., & Kikinis, R. (2012). 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*, *30*(9), 1323-1341. https://doi.org/10.1016/j.mri.2012.05.001

Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., & Yang, X. (2021). A review of deep learning based methods for medical image multi-organ segmentation. *Phys Med*, *85*, 107-122. https://doi.org/10.1016/j.ejmp.2021.05.003

Ge, W., Jing, J., An, S., Herlopian, A., Ng, M., Struck, A. F., Appavu, B., Johnson, E. L., Osman, G., Haider, H. A., Karakis, I., Kim, J. A., Halford, J. J., Dhakar, M. B., Sarkis, R. A., Swisher, C. B., Schmitt, S., Lee, J. W., Tabaeizadeh, M., . . . Brandon Westover, M. (2021). Deep active learning for Interictal Ictal Injury Continuum EEG patterns. *J Neurosci Methods*, *351*, 108966. https://doi.org/10.1016/j.jneumeth.2020.108966

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., Whyntie, T., Nachev, P., Modat, M., Barratt, D. C., Ourselin, S., Cardoso, M. J., & Vercauteren, T. (2018). NiftyNet: a deep-learning platform for medical imaging. *Comput Methods Programs Biomed*, *158*, 113-122. https://doi.org/10.1016/j.cmpb.2018.01.025

Gokila Brindha, P., Kavinraj, M., Manivasakam, P., & Prasanth, P. (2021). Brain tumor detection from MRI images using deep learning techniques. *IOP Conf. Ser.: Mater. Sci. Eng.*, *1055*. https://doi.org/10.1088/1757-899X/1055/1/012115

Grimova, N., & Macas, M. (2019). Query-by-Committee Framework Used for Semi-Automatic Sleep Stages Classification. *MDPI Proceedings*, *31*(80). https://doi.org/10.3390/proceedings2019031080

Hatt, M., Lee, J. A., Schmidtlein, C. R., Naqa, I. E., Caldwell, C., De Bernardi, E., Lu, W., Das, S., Geets, X., Gregoire, V., Jeraj, R., MacManus, M. P., Mawlawi, O. R., Nestle, U., Pugachev, A. B., Schoder, H., Shepherd, T., Spezi, E., Visvikis, D., . . . Kirov, A. S. (2017). Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. *Med Phys*, *44*(6), e1-e42. https://doi.org/10.1002/mp.12124

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. https://doi.org/10.1109/CVPR.2016.90

Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., & Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*, *36*, 61-78. https://doi.org/10.1016/j.media.2016.10.004

Kocher, M., Ruge, M. I., Galldiks, N., & Lohmann, P. (2020). Applications of radiomics and machine learning for radiotherapy of malignant brain tumors. *Strahlenther Onkol*, *196*(10), 856-867. https://doi.org/10.1007/s00066-020-01626-8

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M. J., & Vercauteren, T. (2017). On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain

Parcellation as a Pretext Task. *International Conference on Information Processing in Medical Imaging*, 348-360. https://doi.org/10.1007/978-3-319-59050-9_28

Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., Hawkins, C., Ng, H. K., Pfister, S. M., Reifenberger, G., Soffietti, R., von Deimling, A., & Ellison, D. W. (2021). The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol*, *23*(8), 1231-1251. https://doi.org/10.1093/neuonc/noab106

Madhawa, K., & Murata, T. (2020). Active Learning for Node Classification: An Evaluation. *Entropy (Basel)*, *22*(10). https://doi.org/10.3390/e22101164

Mahapatra, D., Bozorgtabar, B., Thiran, J. P., & Reyes, M. (2018). Efficient Active Learning for Image Classification and Segmentation using a Sample Selection and Conditional Generative Adversarial Network. *arXiv*, arXiv: 1806.05473. https://doi.org/10.48550/ARXIV.1806.05473

Mazzara, G. P., Velthuizen, R. P., Pearlman, J. L., Greenberg, H. M., & Wagner, H. (2004). Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation. *Int J Radiat Oncol Biol Phys*, *59*(1), 300-312. https://doi.org/10.1016/j.ijrobp.2004.01.026

McRobbie, D. W. (2007). *MRI from picture to proton* (2nd ed.). Cambridge University Press. Table of contents only http://www.loc.gov/catdir/enhancements/fy0729/2007277277-t.html

Publisher description http://www.loc.gov/catdir/enhancements/fy0729/2007277277-d.html

Contributor biographical information http://www.loc.gov/catdir/enhancements/fy0733/2007277277-b.html

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M. A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., . . . Van Leemput, K. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging*, *34*(10), 1993-2024. https://doi.org/10.1109/TMI.2014.2377694

Nath, V., Yang, D., Landman, B. A., Xu, D., & Roth, H. R. (2021). Diminishing Uncertainty Within the Training Pool: Active Learning for Medical Image Segmentation. *IEEE Trans Med Imaging*, *40*(10), 2534-2547. https://doi.org/10.1109/TMI.2020.3048055

Neugut, A. I., Sackstein, P., Hillyer, G. C., Jacobson, J. S., Bruce, J., Lassman, A. B., & Stieg, P. A. (2019). Magnetic Resonance Imaging-Based Screening for Asymptomatic Brain Tumors: A Review. *Oncologist*, *24*(3), 375-384. https://doi.org/10.1634/theoncologist.2018-0177

Nock, R., & Nielsen, F. (2004). Statistical region merging. *IEEE Trans Pattern Anal Mach Intell*, *26*(11), 1452-1458. https://doi.org/10.1109/TPAMI.2004.110

Odland, A., Server, A., Saxhaug, C., Breivik, B., Groote, R., Vardal, J., Larsson, C., & Bjornerud, A. (2015). Volumetric glioma quantification: comparison of manual and semi-automatic tumor segmentation for the quantification of tumor growth. *Acta Radiol*, *56*(11), 1396-1403. https://doi.org/10.1177/0284185114554822

Ostrom, Q. T., Bauchet, L., Davis, F. G., Deltour, I., Fisher, J. L., Langer, C. E., Pekmezci, M., Schwartzbaum, J. A., Turner, M. C., Walsh, K. M., Wrensch, M. R., & Barnholtz-Sloan, J. S. (2014). The epidemiology of glioma in adults: a "state of the science" review. *Neuro Oncol*, *16*(7), 896-913. https://doi.org/10.1093/neuonc/nou087

Ostrom, Q. T., Patil, N., Cioffi, G., Waite, K., Kruchko, C., & Barnholtz-Sloan, J. S. (2020). CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013-2017. *Neuro Oncol*, *22*(12 Suppl 2), iv1-iv96. https://doi.org/10.1093/neuonc/noaa200

Pereira, S., Pinto, A., Alves, V., & Silva, C. A. (2016). Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Transactions on Medical Imaging*, *35*(5), 1240-1251. https://doi.org/10.1109/TMI.2016.2538465

Pham, D. L., Xu, C., & Prince, J. L. (2000). Current methods in medical image segmentation. *Annu Rev Biomed Eng*, *2*, 315-337. https://doi.org/10.1146/annurev.bioeng.2.1.315

Qian, P., Chen, Y., Kuo, J. W., Zhang, Y. D., Jiang, Y., Zhao, K., Al Helo, R., Friel, H., Baydoun, A., Zhou, F., Heo, J. U., Avril, N., Herrmann, K., Ellis, R., Traughber, B., Jones, R. S., Wang, S., Su, K. H., & Muzic, R. F. (2020). mDixon-Based Synthetic CT Generation for PET Attenuation Correction on Abdomen and Pelvis Jointly Using Transfer Fuzzy Clustering and Active Learning-Based Classification. *IEEE Trans Med Imaging*, *39*(4), 819-832. https://doi.org/10.1109/TMI.2019.2935916

Rehman, M. U., Cho, S., Kim, J., & Chong, K. T. (2021). BrainSeg-Net: Brain Tumor MR Image Segmentation via Enhanced Encoder-Decoder Network. *Diagnostics (Basel)*, *11*(2). https://doi.org/10.3390/diagnostics11020169

Rohlfing, T., Zahr, N. M., Sullivan, E. V., & Pfefferbaum, A. (2010). The SRI24 multichannel atlas of normal adult human brain structure. *Hum Brain Mapp*, *31*(5), 798-819. https://doi.org/10.1002/hbm.20906

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev*, *65*(6), 386-408. https://doi.org/10.1037/h0042519

Shapiro, L. G., & Stockman, G. C. (2001). *Computer vision*. Prentice Hall.

Sourati, J., Gholipour, A., Dy, J. G., Kurugol, S., & Warfield, S. K. (2018). Active Deep Learning with Fisher Information for Patch-wise Semantic Segmentation. *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018)*, *11045*, 83-91. https://doi.org/10.1007/978-3-030-00889-5_10

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*, *15*, 29. https://doi.org/10.1186/s12880-015-0068-x

Top, A., Hamarneh, G., & Abugharbieh, R. (2011). Active learning for interactive 3D image segmentation. *Med Image Comput Comput Assist Interv*, *14*(Pt 3), 603-610. https://doi.org/10.1007/978-3-642-23626-6_74

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*, *29*(6), 1310-1320. https://doi.org/10.1109/TMI.2010.2046908

Wang, W., Hu, Y., Lu, P., Li, Y., Chen, Y., Tian, M., & Yu, L. (2014). Evaluation of the diagnostic performance of magnetic resonance spectroscopy in brain tumors: a systematic review and meta-analysis. *PLoS One*, *9*(11), e112577. https://doi.org/10.1371/journal.pone.0112577

Wanis, H. A., Moller, H., Ashkan, K., & Davies, E. A. (2021). The incidence of major subtypes of primary brain tumors in adults in England 1995-2017. *Neuro Oncol*, *23*(8), 1371-1382. https://doi.org/10.1093/neuonc/noab076

WHO Classification of Tumours Editorial Board. (2021). *Central Nervous System Tumours* (5th ed., Vol. 6). International Agency for Research on Cancer.

Wu, W., Chen, A. Y., Zhao, L., & Corso, J. J. (2014). Brain tumor detection and segmentation in a CRF (conditional random fields) framework with pixel-pairwise affinity and superpixel-level features. *Int J Comput Assist Radiol Surg*, *9*(2), 241-253. https://doi.org/10.1007/s11548-013-0922-7

Yang, T., Zhou, Y., Li, L., & Zhu, C. (2020). DCU-Net: Multi-scale U-Net for brain tumor segmentation. *J Xray Sci Technol*, *28*(4), 709-726. https://doi.org/10.3233/XST-200650

Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., & Palmer, A. C. (1994). Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans Med Imaging*, *13*(4), 716-724. https://doi.org/10.1109/42.363096