uni.lu

UNIVERSITÉ DU LUXEMBOURG

Faculty of Science, Technology, and Medicine

Master in Integrated Systems Biology

MASTER Thesis

by

Gonçalo Artur Gaspar Alves

Born on 1st January 1999 in Lisbon (Portugal)

On the distribution of gliomas in publicly available MRI datasets and the recognition of duplicate images

UNIVERSITÉ DU LUXEMBOURG

Faculty of Science, Technology, and Medicine

Master in Integrated Systems Biology

MASTER Thesis

by

Gonçalo Artur Gaspar Alves

Born on 1st January 1999 in Lisbon (Portugal)

On the distribution of gliomas in publicly available MRI datasets and the recognition of duplicate images

Defence:	16 July 2024 in Luxembourg
Supervisors:	Andreas Husch, Principal Investigator, LCSB (Luxembourg) Ben Bausch, Doctoral Researcher, LCSB (Luxembourg)
Jury members:	La Mi, Postdoctoral researcher, LCSB (Luxembourg) Beatriz García Santa Cruz, Postdoctoral researcher, LCSB (Luxembourg)

Information on thesis

This Thesis was written at the Luxembourg Centre for Systems Biomedicine in the ImagingAI group under the supervision of Dr. Andreas Husch, Dr. François Lamoline, and Ben Bausch.

The Luxembourg Centre for Systems Biomedicine (LCSB) is a multidisciplinary institute, drawing from several biological and computational disciplines, focusing on research pertaining to the brain and its diseases. Specifically, this institute is focused on neurodegenerative diseases such as Alzheimer's disease or Parkinson's disease.

Integrated within the LCSB, the Imaging AI (IAI) group, endeavors to translate recent discoveries and breakthrough in the field of Artificial Intelligence (AI), biomedical signal analysis, and medical image computing, into the field of biology/medicine to solve concrete problems within these fields.

Dr. Andreas Husch is a research scientist and principal investigator of the IAI group. His primary research subject is computer science, and his research topics of interest are Signal and Image Processing, Image Guided Interventions, Applied Machine Learning, and Computational Neuroscience.

Dr. François Lamoline is a postdoctoral researcher within the IAI group. His primary research subject is mathematics, and his research topics of interest are; Distributed-parameter Systems, System Identification, Stochastic Modelling, Multi-modal Learning, and Port Hamiltonian Framework

Ben Bausch is a doctoral researcher within the IAI group. His primary research subjects are Explainable Deep Neural Networks for Bioimaging, and Computervision.

Table of contents

List of Figures

Abstract

The rapid evolution of Deep Learning (DL) models has revolutionized numerous domains, including medical imaging. However, the philosophy of ”move fast and break things” can be problematic in medical contexts, where technological advancements have profound impacts on patient care. This thesis investigates two pivotal aspects of publicly available MRI datasets for gliomas, tumor distribution and dataset curation.

Firstly, the study focuses on the statistical distribution of gliomas within MRI datasets, addressing the variability in tumor distribution across different brain regions. It evaluates the effectiveness of two registration techniques, rigid and deformable registration, in aligning MRI data with brain atlases. The research demonstrates that deformable registration offers superior alignment, reducing misalignment. Furthermore, analysis of the tumor distribution in several datasets shows differential tumor distribution in these datasets. This insight is crucial to inform researchers of the when selecting a dataset for research purposes.

Secondly, the thesis presents Tensor-based Analysis for Redundant Knowledge in Neuroimaging (TARKIN), a novel DL model developed to identify duplicate MRI images within self-curated datasets. TARKIN utilizes a modified UNET architecture to detect duplicates effectively. The model's performance was assessed using various embedding, reconstruction, and combination losses. Furthermore, different DL architectures were assessed to solve this task. Results indicate that TARKIN excels in identifying duplicates in non-augmented images but faces challenges with augmented data, particularly images subjected to canonical transformations. The findings reveal that while reconstruction losses are beneficial for handling light augmentations, a combination of losses may enhance performance with more complex data augmentations.

This work contributes to the field by improving the understanding of glioma distribution in MRI datasets and providing a practical tool for duplicate image detection. The insights gained from this research are valuable for researchers and practitioners in selecting and curating datasets for deep learning applications, ultimately supporting the development of more accurate and reliable medical imaging models. This thesis brings the importance of data curation to the foreground and offers a solution to enhance the quality and usability of medical imaging datasets.

Acronyms

ADAM Adaptive Moment Estimation. 18

AI Artificial Intelligence. i, 9

ANTS Advanced Normalization Tool. 23

BIDS Brain Imaging Data Structure. 21

BraTS-2020 Brain Tumor Segmentation 2020. 43, 51, 52, 56

BraTS-2021 Brain Tumor Segmentation 2021. iv, v, 21, 22, 29–31, 34, 36, 43, 46, 51, 52, 54, 55, 79

BraTS-SSA Brain Tumor Segmentation Sub-Saharan Africa. iv, v, 21, 22, 30–32, 34–36, 54, 55, 80

BU-NET Basic U-NET. 42, 48

Burdenko-GBM Burdenko Glioblastoma Progression. iv, v, 22, 25, 31, 34–36, 54, 81

CaPTk Cancer Imaging Phenomics Toolkit. 21–23, 53

CNN Convolutional Neural Network. 2, 15, 18

DL Deep Learning. vi, 1, 2, 9–11, 19, 20, 23, 37, 43, 56, 57, 71

DWI Diffusion Weighted Imaging. 27, 56

HPC High-performance Computer. 21, 23, 53

IAI Imaging AI. i, 21

ICBM International Consortium of Brain Mapping. 6

LCSB Luxembourg Centre for Systems Biomedicine. i, 21

LGG Low Grade Gliomas 1p19q Deletion. iv, v, 21, 22, 29–31, 34–36, 54–56

ML Machine Learning. 9

MNI Montreal Neurological Institute. 6

MONAI Medical Open Network for Artificial Intelligence. 41, 42, 48, 57

MRI Magnetic Resonance Imaging. iv, vi, 1, 2, 4, 6, 20, 21, 23–26, 28, 29, 37, 38, 43, 46, 53, 54, 56–58, 60, 70

MU-NET Modified U-NET. 41, 42, 48, 49, 51

NCC Normalized cross-correlation. 23

NMI Normalized mutual information. 23, 24

QIN Quantitative Imaging Network. iv, v, 22, 31, 34–36, 54, 55, 82

RHUH-GBM Río Hortega University Hospital Glioblastoma. v, 21, 22, 31, 34–36, 54, 83

SNN Siamese Neural Network. 9, 20, 37, 38, 41, 49, 57

SRI Stanford Research Institute. 6

SSD Sum of squared differences. 23

SSIM Structural Similarity Index Measure. 46, 47

TARKIN Tensor-based Analysis for Redundant Knowledge in Neuroimaging. vi, 60

TCIA The Cancer Imaging Archive. 21

UCSF-PDGM University of California San Francisco Preoperative Diffuse Glioma. v, 22, 30, 31, 34, 36, 54, 55, 84

URT Unified Retrieval Tool. 21

1 Introduction & Background

The fast-paced evolution of DL models in our current age is unprecedented. The utility of these models is undeniable, however the philosophy of "move fast and break things" might be misplaced, especially in the medical domain in which technological developments can have real, positive and/or negative, effects on the most vulnerable amongst us. For this reason, it becomes important to ascertain and curate the quality of one of the most important components on which DL models rely: the data.

The training of DL models for medical purposes is dependent on access to a dataset of medical images to train these models on. The datasets, and the quality thereof, affect the performance of a DL model, the general rule of thumb being that large datasets with many images result in models that are more performant than smaller datasets with fewer images. However, in medical imaging, large datasets are difficult to come by due to a variety of reasons, such as the cost associated to taking MRIs, the time required to create a dataset of substantial size, and the convoluted nature of MRI formatting [1]. Hence, any task to be undertaken using any of the openly available medical image datasets should be conducted using the most extensive information available on these datasets.

This thesis strives to supplement two areas of publicly available MRI datasets containing gliomas, usually used in tasks such as tumor segmentation, augmentation, and survival prediction, that are lacking in the literature. Firstly, it endeavors to determine the statistical distribution of gliomas, the brain regions and structures that are primarily affected by gliomas, and the attainment of the connectome in several open-source datasets. Secondly, it proposes a DL model, based on a modified UNET and a Siamese Neural Network, that is able to detect duplicate MRIs in self-curated datasets. Furthermore, it seeks to ascertain what losses are most indicated for detection of MRI duplicates. The overarching objective being that both of these tasks will positively contribute to the curation and maintenance of MRI datasets, and inform researchers of the statistical characteristics of these datasets to inform their decision when selecting an open-source dataset for training DL models.

Extensive information is available on various aspects of MRI datasets containing gliomas, typically limited to basic dataset features such as patient demographics, tumor counts, and image characteristics (resolution, data type, voxel size, etc.). However, analyses focusing on tumor distribution within these datasets are often missing. Additionally, recent research highlights an intricate relationship between the connectome and gliomas, which is also not considered in these datasets [2][3]. These factors can be especially important in an age where DL, and its reliance on data distributions, play an ever larger role in the global MRI research landscape.

Creating and curating new MRI datasets involves certain risks. MRIs are notoriously scarce, and DL methodologies are notoriously data-hungry. Hence a solution to this will inevitably be the merging of several independent MRI datasets to create a large dataset on which to train a DL model. However, this approach risks including duplicateMRIs, a problem that becomes more challenging as MRIs are shared, modified, and shared again. To mitigate this, a specialized tool is needed to detect MRI duplicates.

This thesis is organized as follows. The first section of this thesis covers the necessary background to understand the methods utilized in this work. It primarily focuses on theoretical concepts important to understand a Convolutional Neural Network (CNN), medical imaging and MRI, and various brain health concepts, including gliomas and the connectome. The second section details the specific methods used in this thesis and the results they yielded. Lastly, the third section is dedicated to discussion and future outlook.

1.1 Medical Imaging & Gliomas

This section covers several concepts that are important to understand the work performed within this thesis. It covers the general concepts in neuroimaging and MRI, defines gliomas and their perception in MRI, glioma migration and invasion of healthy brain tissue and its link to the connectome.

1.1.1 Neuroimaging

Neuroimaging refers to a branch of medical imaging that focuses on diagnosing brain related diseases, and ascertain brain health. It also studies how the brain works and how certain activities affect brain function. There are two broad categories of neuroimaging: structural imaging, which quantifies brain structure, and functional imaging, which studies brain function. Common techniques include MRI, positron emission tomography (PET), and functional magnetic resonance imaging (fMRI). For the purposes of this report, only MRI images are relevant [4].

Neuroimaging has become increasingly relevant to study brain afflictions. In the past, studying this kind of disease was only possible through autopsies, or by observing the symptoms of the patient. However, with the surge of advancements in the field of neuroimaging it has become possible to elucidate on the brain's chemistry, physiology, and electrical and metabolic activity in real time. Furthermore, this field is becoming increasingly important in Luxembourg as the University of Luxembourg is increasingly invested in researching neurodegenerative diseases such as forms of dementia, and especially Alzheimer's disease [5].

1.1.2 Magnetic Resonance Imaging (MRI)

An MRI is a type of medical imaging procedure that outputs an image of the internal structures of the body. Unlike other methods of medical imaging such as simple X-rays or Computed Tomography (CT) scans, MRIs do not produce any ionizing radiation, rather they produce a magnetic field that permits the image taking [6]. MRI makes use of the fact that different tissues interact differently with electromagnetic waves. During an MRI procedure, a strong magnetic field is generated, which aligns the protons in water to the magnetic field. A second electromagnetic wave is generated to disrupt the alignment of the protons caused by the electromagnet field. As this disruption is corrected, and the protons return to their initial position of alignment with the magnetic field, an electromagnetic wave is generated which can be detected by the detectors in the MRI machine. As mentioned before, different tissues interact differently with this wave and will appear differently in the final image [7].

The most common types of images seen when looking at MRI images are T1 and T2 images. As mentioned before, during the MRI procedure, a magnetic field aligns the protons to the direction of the magnetic field. This created a longitudinal magnetization vector. You can picture this magnetization vector as a vector of a certain magnitude pointing in the direction of the y-axis on a 2D graph. When the second electromagnetic pulse is applied momentarily, usually in the form of a Radiofrequency (RF) pulse, the longitudinal magnetization vector decays into a horizontal magnetization vector, which you can picture as a vector of a certain magnitude pointing in the direction of the x-axis in a 2D graph. Once the RF pulse stops, the horizontal magnetization vector will begin to decay. The T2 image is measured at this stage until 63% of the horizontal magnetization vector has decayed, which is termed as T2 time. Whilst the horizontal magnetization vector decays, the longitudinal magnetization vector grows. The T1 image is measured in this stage until 63% of the longitudinal magnetization vector was recovered, which is termed as T1 time. The T1 and T2 times are unique to each tissue, which explains how different structures can be distinguished in MRI images [7].

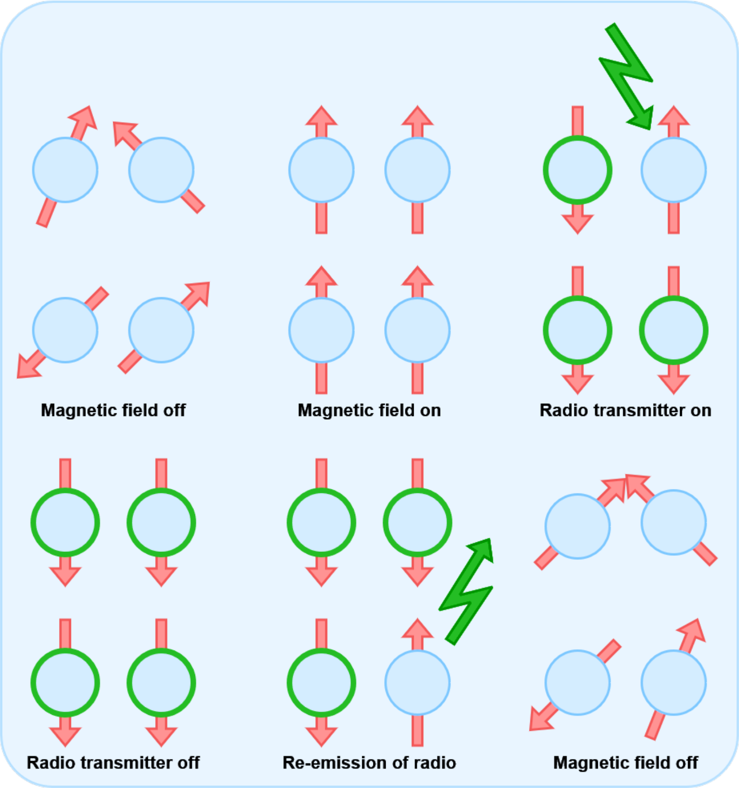


Figure 1: Alignment of protons during an MRI procedure. (Top left) Protons are unaligned when there is no electromagnetic field. (Top Middle) Protons are aligned once an electromagnetic field is induced. This causes the longitudinal magnetization vector. (Top Right) An RF pulse is shortly applied, which causes the protons to align differently. (Bottom Left) The change of proton alignment causes the horizontal magnetization vector. (Bottom Middle) The emission of the electromagnetic wave that is generated whilst the proton returns to alignment with the electromagnetic field, which is at a lower energy level, is picked up by the detector of the MRI machine. (Bottom Right) The protons return to their normal state once the electromagnetic field is turned off.

Another important characteristic of MRIs is that the electromagnetic signals are usually given by protons contained in water, meaning that structures that have very little water give off very small signals and can be barely seen in MRIs, such as the lungs.

MRIs can be affected by several artifacts at the moment the image is being captured. These artifacts can stem from several different sources, such as; motion of the patient, ornamentation of the body (tattoos, make-up, etc.), medical devices and interventions (hearing aids, prostheses, dental implants, ligament reconstruction), and from the MRI machine itself (bias field artifacts, spike artifacts, noise) [8][9]. In addition to these artifacts, or augmentations, created during the generation of the image, augmentations of MRI images can occur as the images pass from research group to research group, or are disseminated over the internet. These augmentations would be flipping the image causing a laterality error, or poor treatment of the data that can cause voxel anisotropy and introducing noise.

1.1.3 MRI modalities

MRI modalities refer to the different types of MRI scans that can be performed to visualize various aspects of the body's anatomy and function. There is a wide variety of MRI modalities, the most common of which are the T1 and T2 modalities. For this report, FOUR modalities were of great importance [10]:

- **T2**: T2 images are valuable for observing fluids, since this modality enhances the water signal. It is useful for observing structures, or pathologies, that contain a lot of fluid, such as edema, and delineating between different types of soft tissues. For the rest of the structures, the intensity of the signal depends on the T2 properties of the tissue

- **T1**: In contrast to T2, T1 enhances the signal of fatty tissues and suppresses that of fluids. For the rest of the structures, the intensity of the signal depends on the T1 properties of the tissue.

- **T1 contrast enhanced (T1ce)**: This modality still makes use of a T1 scan, however, it includes the administration of a contrast agent, typically gadolinium. It is particularly useful for visualizing blood-brain barrier disruption and enhancing certain types of lesions, such as tumors

- **Fluid Attenuated Inversion Recovery (FLAIR)**: FLAIR are commonly used in brain imaging, and suppress the signal from the cerebrospinal fluid (CSF), while enhancing the signal from lesions or pathological processes, such as infections, demyelination, or tumors.

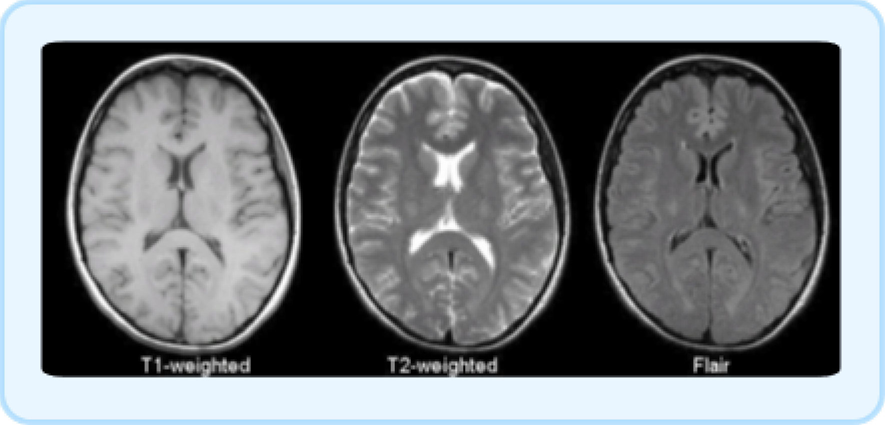


Figure 2: Example of different MRI modalities. (Left) T1-weighted image. The cerebrospinal fluid is not visible. The image is also generally clearer than a T2 image due to the fatty nature of the brain. (Middle) T2-weighted image. The cerebrospinal fluid gives a strong signal and the brain is darker than in the T1. (Right) FLAIR image. The signals from fluids are reduced, and there are no noticeable signals, since this brain has no lesions.

1.1.4 MNI atlas & standard spaces

A ”standard space”, in the context of MRI and medical imagery, refers to defined boundaries around an organ, expressed in millimeters, from a set origin. This term is used interchangeably alongside ”brain template” and ”brain atlas”. They are employed to refer to an average, populational, depiction of the brain, as can be seen in *Figure 3*. For simplicity, in this thesis, a ”brain atlas” will refer to an average depiction of the brain, and a ”brain template” will refer to a specific iteration or version of a brain atlas.

The MNI152 atlas was created by the Montreal Neurological Institute (MNI) [11]. It represents an attempt to create an average depiction of the human brain. It is an average of 152 T1-weighted MRI scans from different patients, linearly transformed to a previously popular brain atlas called Talairach. The MNI152 was adopted to define standard anatomy by the International Consortium of Brain Mapping (ICBM). There are several templates of the MNI152 atlas that are routinely employed, and also different atlases, as mentioned previously. For this thesis, the most important atlases are; the MNI152 atlas, and the Stanford Research Institute (SRI) [12]. Important templates are; MNI152 2009a NLIN asymmetric T1, MNI152 2009c NLIN asymmetric T1, SRI24, and the Harvard-Oxford cortical and subcortical structural templates.

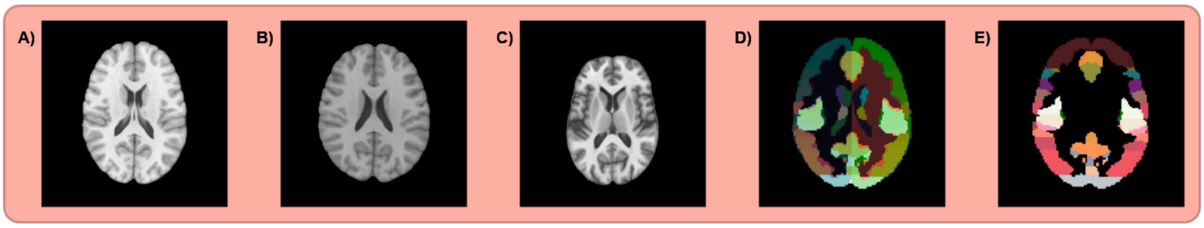


Figure 3: The brain atlases and templates used in this thesis. (A) MNI152 2009a NLIN asymmetric T1, (B) MNI152 2009c NLIN asymmetric T1, (C) SRI24, (D) MNI152 Harvard-Oxford cortical, (E) MNI152 Harvard-Oxford subcortical.

1.1.5 Glioma

Gliomas are a general category that encompasses brain tumors that emerge from glial cells[13]. Gliomas account for about 30% of all brain tumors, but also account for 80% of all malignant brain tumors [14]. These highly vascular tumors can easily infiltrate different tissues and spread, due to these characteristics they often reappear after surgery or treatment. This type of tumor is often a focus in research groups due to their preponderance of malignancy and due to their high vascularity [15].

The symptoms of gliomas are generally headaches, vomiting, and/or seizures caused by the increased intracranial pressure. However, they can be different depending on the region that is affected [16]. For instance, a glioma affecting the optic nerve or visual cortex can affect vision and lead to vision loss [17].

The causes that lead to the development of gliomas are varied. Many mechanisms and causes have been suggested, and they vary depending on the type of glioma. A well-known and documented risk seems to be radiation. Exposure to ionizing radiation such as that obtained during a CT scan can increase the risk of developing a glioma by up to 55% [18]. Gliomas have also been positively associated with an infection of Cytomegalovirus [19]. Other reasons include Diet, inherited polymorphisms of DNA repair genes [20], and hereditary disorders [21].

Given that glioma is an umbrella term for a wide category of tumors, a glioma can be classified by 3 categories: by type of cell, by location, and by grade [22]. The cell type refers to the type of cell with which the glioma shares the most histological resemblance, and not necessarily to the cell type from which it originated. For instance, an astrocytoma, means that this specific gliomas shares its histological features with astrocytes, but it could have originated from an oligodendrocyte. The classification by location refers to if the glioma is above or below the tentorium. Gliomas that are above the tentorium are classified as supratentorial and affect the cerebrum, and gliomas below the tentorium are infratentorial and affect the cerebellum. The classification by grade refers to the pathological evaluation of the glioma. Biologically benign gliomas are small and can be removed through surgery. Low-grade gliomas are not anaplastic, well differentiated tumors that are low-risk but can turn malignant eventually. This type of tumor is often monitored if it can not be removed, and if it does not hinder the patient. Patients can live many years with this grade of tumor. High-grade gliomas are anaplastic, undifferentiated tumors and present the greatest danger for the patient.

Gliomas are characterized for exhibiting 3 regions. Firstly, the edema: a swelling caused by the accumulation of fluid in the surrounding brain tissue. Secondly, the enhancing region: the part of the tumor that is actively growing. Finally, the necrotic core: the central area of the tumor that has undergone necrosis, meaning the tumor cells in this region have died due to lack of blood supply or other factors

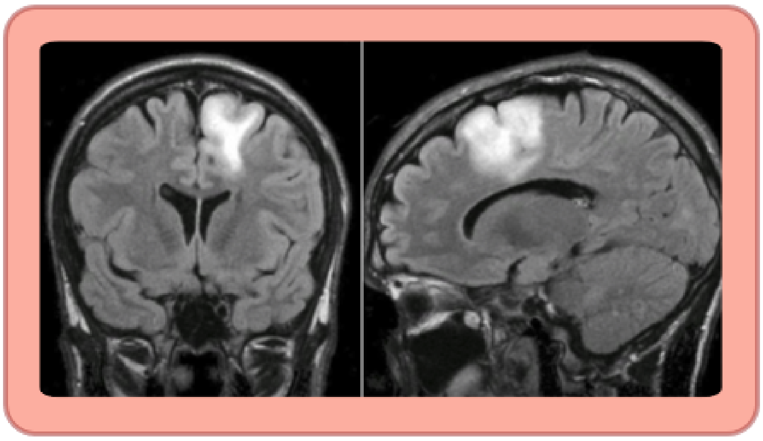


Figure 4: Example of a Glioma lesion in MRI. (Left) Glioma in a FLAIR coronal MRI. (Right) Glioma in a FLAIR sagittal MRI.

1.1.6 Connectome and cancer

The connectome refers to the hierarchical networks that explain the structural and functional mappings of the brain. When first introduced in 2005 by Sporns et al., the connectome advanced the understanding of how brain function arises from its base structure, and explained the dynamic interaction between brain function and structure[23]. For the scope and context of this thesis, it is sufficient to think of the connectome as a map of white matter tracts and cortical functional connections. Recently, the connectome has been shown to be involved in carcinogenesis. Pathologies, such as those involving misfolded proteins and cancer, have been shown to spread preferentially through anatomical pathways, indicating that these pathologies make use of the connectome and its vast array of white matter tracts. In fact, some mechanisms have been proposed that associate cancer migration and invasion with several axonal guidance proteins [24].

This association between white matter tracts and glioma migration has also been shown to affect patient survivability [2]. Patients that had glioma in regions with a high tract density index, had a lower survival rate than those that had glioma in regions with a low tract density index. Which indicates that not only location but also interference and attainment of white matter tracts is an important indicator of patient survivability.

1.2 Machine Learning

This section provides a definition of DL, Neural Networks, and several Neural Network architectures, namely Siamese Neural Network (SNN)s, and U-NET. Furthermore, it also provides a mathematical definition of the losses that were used during the DL experiments within the context of this thesis.

1.2.1 Deep Learning

Machine Learning (ML) is a field of research in AI that seeks to study and develop algorithms that can learn patterns in data that is shown to them, without getting specific instructions, and generalize what it has learned to data that it has not yet seen, *Figure A.2.* ML methods conventionally range from simple algorithms like linear regressions in high dimensional data, to complex methods like deep neural networks. However, traditional methods they rely on heavy human intervention and handcrafted features to ensure that the algorithms would extract the right information from the data and transform this raw data into a suitable internal representation. Subsequently, an internal system would use the generated internal representation to solve a wide range of different tasks.

DL methods make use of representation learning, which allows for a machine to be shown raw data and, by itself, find a suitable representation of that data in vector space. DL allows the machine to have several layers of representation at different scales, connected together in a non-linear way with each layer transforming the original data in different ways, and consequently, several layers of abstraction. These transformations allow the machine to learn very complex functions. Distinctively, the representations obtained from these layers are not designed by humans, they are learned by the machine during training [25].

1.2.2 Fully Connected Layer

Traditionally, neural networks were constructed as fully connected layers, also referred to as dense layers. In a fully connected layer, each input feature is connected to each output feature in a linear fashion by a function. Depictions can be found in the appendix *Figure A.3*, *Equation 12*. Each connection in the fully connected layer is a learnable parameter. The size of the fully connected layer is a set of nodes that is as large as the user's context needs it to be. For instance, for a classification task, in which there are two classes, the fully connected layer at the end of the network would have two output nodes each corresponding to one of the classes.

1.2.3 Convolutional Layer

A convolution is a type of linear operation where a small matrix, usually 3x3, called a "kernel" is used to extract features from an input matrix, usually called a "tensor", such as an image. An element-wise product is calculated from the kernel and the tensor and the output is written to the corresponding position in an output tensor, usually called a feature map. This element-wise product is done for every location on the input tensor. This process, depicted in *Figure 5* can be applied an arbitrary amount of times to the same input tensor with different kernels, which results in different feature maps, also called channels.

The general equation for the convolution operation of a 2D image is given below,

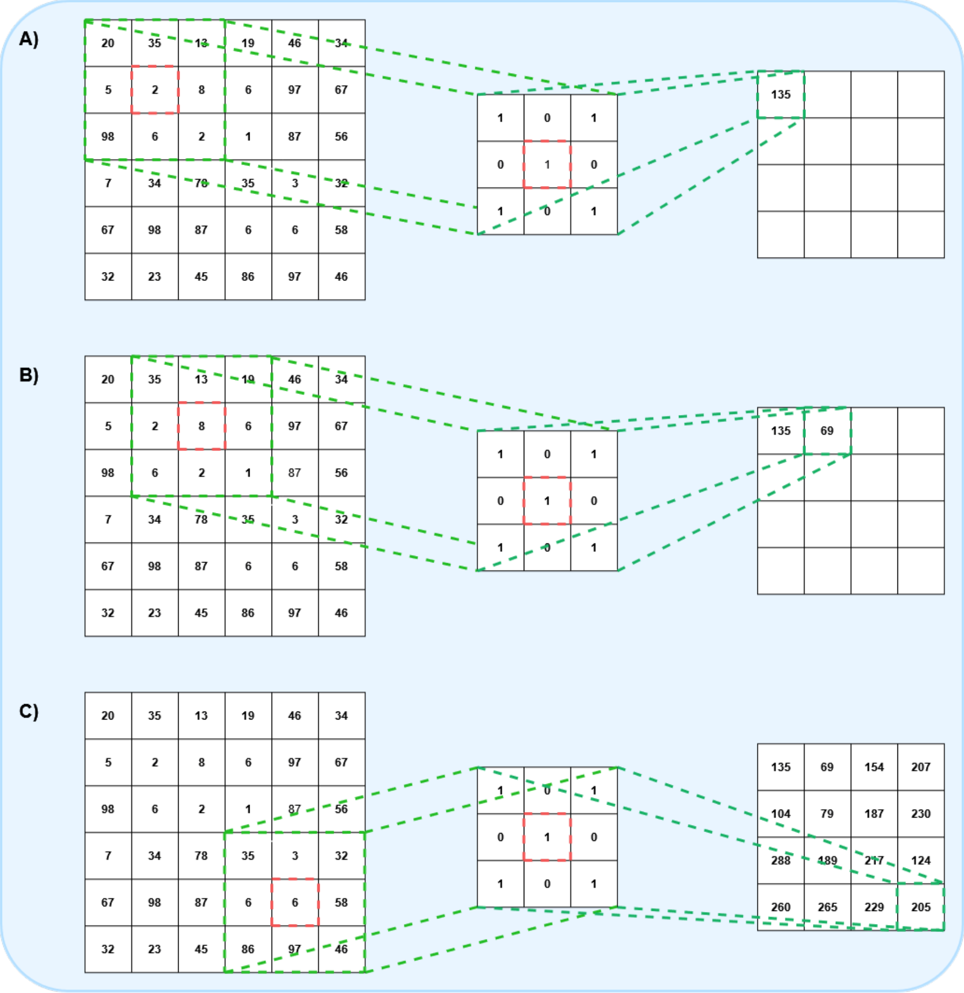
$$(f * g)[i,j] = \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} f[i-m, j-n] \cdot g[m,n], \tag{1}$$

where:

- $f[i,j]$ represents the 2D input signal,
- $g[m,n]$ represents the 2D convolution kernel,
- i, j are the coordinates of the output pixel,
- m, n are the coordinates in the kernel

In DL, a hyperparameter is a configuration that is external to the model. They are tuned to a specific problem, and set by the researcher. A parameter is a variable that is internal to the model. Parameters are adjusted and learned from the data, and can't be set by the researcher [26]. The size of the kernel is a hyperparameter. As mentioned before, it is usually

3x3, however, this can be set by the user to correspond to the context of his needs. Similarly, the number of feature maps that are extracted from an input tensor is also a hyperparameter and can be set by the user to fit his specific context.



A)

20	35	13	19	46	34
5	2	8	6	97	67
98	6	2	1	87	56
7	34	78	35	3	32
67	98	87	6	6	58
32	23	45	86	97	46

1	0	1
0	1	0
1	0	1

135			

B)

20	35	13	19	46	34
5	2	8	6	97	67
98	6	2	1	87	56
7	34	78	35	3	32
67	98	87	6	6	58
32	23	45	86	97	46

1	0	1
0	1	0
1	0	1

135	69		

C)

20	35	13	19	46	34
5	2	8	6	97	67
98	6	2	1	87	56
7	34	78	35	3	32
67	98	87	6	6	58
32	23	45	86	97	46

1	0	1
0	1	0
1	0	1

135	69	154	207
104	79	187	230
288	189	217	124
260	265	229	205

Figure 5: A convolution operation using a 3x3 kernel on an input tensor of size 6x6 with a stride of 1, and no padding leading to a downsampling of the image, and a corresponding output tensor of size 4x4. An element-wise product is calculated for each element of the kernel and input tensor. The sum of all of these products at one specific location is written to the corresponding position on the output tensor. (A) The initial position of the kernel on the input tensor, the elements considered for the element-wise product, and the corresponding output in the downsized output tensor. (B) The kernel moving one position to the right, due to the stride being 1, and performing the same operation to obtain the next value for the next position of the output tensor. (C) The kernel iterates over the whole input tensor.

Differing hyperparameters affect the resulting output tensor from a convolution operation. For instance, *Figure 5*, shows a convolution operation with no padding. Meaning that the resulting feature map will be downsized when compared to the input tensor. This can be a problem in some DL model architectures when concatenation of two related tensors is required after a convolution. For this reason, some convolutional operations add a layer of padding, either 0-value padding or border padding, around the input tensor to conserve its shape. Furthermore, the stride can be changed according to the necessity of the user. In *Figure 5*, the stride is 1, meaning that the kernel will move by 1 position in the grid relative to its central box. A stride of 2 would have the kernel move 2 positions, and so on. The stride affects the degree of downsampling of the image, with higher strides leading to greater downsampling.

Although the stride can be used to downsample an input tensor, this is usually not the case in convolutional layers. Convolutional layers rely on a further operation after the convolution operation called the pooling operation, often also referred to as the pooling layer. The pooling layer is charged with downsampling the input tensor before it is subjected to the next convolution operation. This has two major advantages. By downsampling the input tensor, an invariance to translation for small shifts and distortions is introduced, and the number of learnable parameters is reduced. In contrast to the convolution operation, the kernel of a pooling operation is not a learnable parameter [27].

The most common pooling operation is max pooling. In max pooling, a kernel of size 2x2 with a stride of 2 moves across an input tensor. At each step of the kernel, the largest local number is kept, and the rest are discarded, as is exemplified in *Figure 6.*

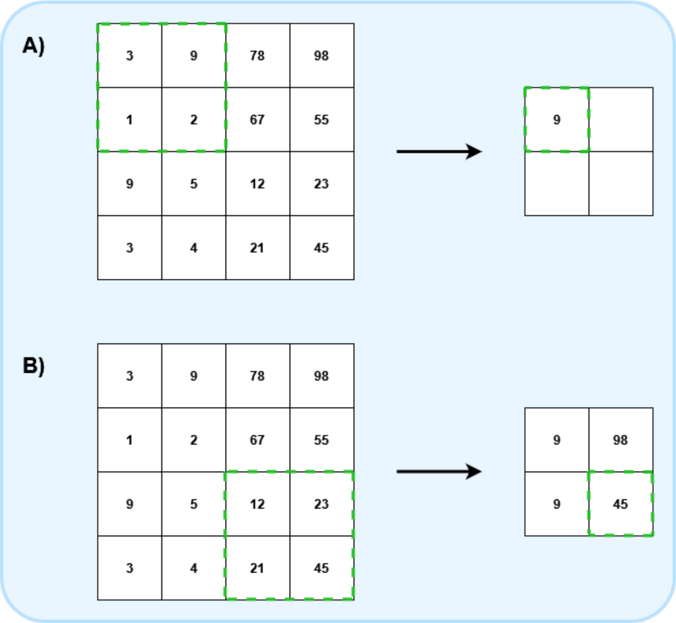


Figure 6: The basic concept of a pooling operation. In this case of a max pooling. For a kernel size of 2x2 and a stride of 2, the local maximum value is retained for the downsized feature map, and the remaining values are discarded. (A) The initial step of a max pooling operation. (B) The final step of a max pooling operation.

A convolution operation, followed by a pooling operation, and an activation function is referred to as a convolutional layer.

An important caveat of the convolutional layer, is that the weights of the kernels are shared across each position of the kernel over an input tensor. This gives the convolutional layer several valuable features. For instance, since the same kernel is being passed over every position of the image, the feature patterns that the kernel is able to detect become invariant to translation. The same pattern will be recognized over the whole image. Furthermore, in conjunction with the pooling operation, the deeper the network goes, the more complex the features it can extract get, and the wider its field of view gets. The initial layers might detect simple patterns like edges, while deeper layers can capture more complex structures like shapes or even objects. This hierarchy helps the network to build a detailed understanding of the image content from simple to complex features, with relatively few parameters.

1.2.4 Activation Functions

Each convolutional layer, and more generally each layer in a neural network, is followed by an activation function. Usually, the role of the activation function is to introduce non-linearity to the network. This counteracts the fact that the linear combination of linear operations is a linear function. This step makes it possible for the network to learn non-linear relationships in data. Several activation functions can be used depending on the task at hand. *Figure 7* contains a few commonly used activation functions.

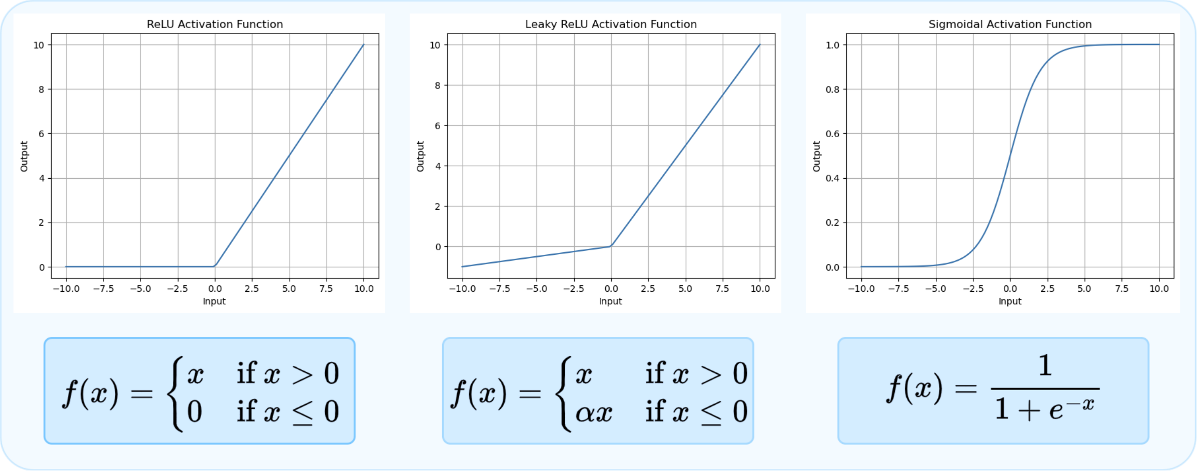
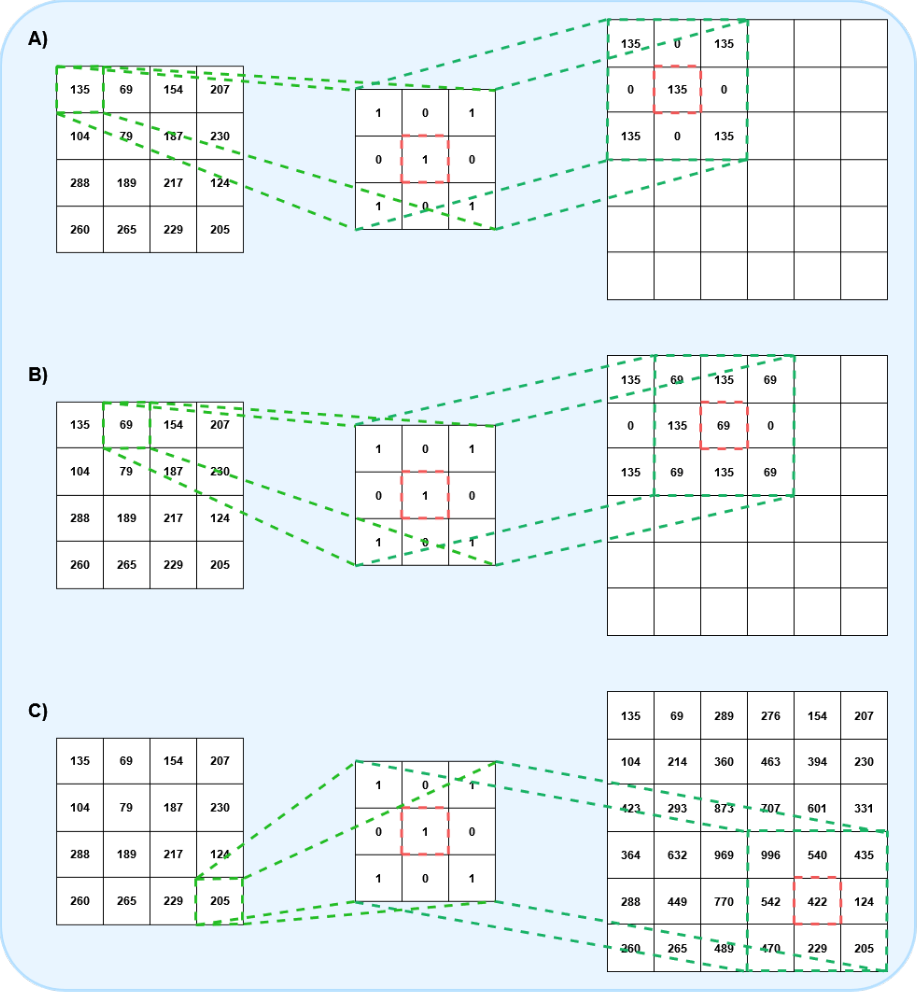


Figure 7: Commonly used activation functions. (Top Left) Rectified Linear Unit (ReLU). (Top Middle) Leaky Rectified Linear Unit (Leaky ReLU). (Top Right) Sigmoid. (Below) The accompanying analytical form of each activation form.

The last layer of a neural network also makes use of an activation function. However, the goal is not to introduce non-linearity into the network. Depending on the task at hand the activation function would be different, and have a different role. For instance, a softmax activation function would be used for a multi-class classification task, because the output of this function would lie between 0 and 1, and sum up to 1. This characteristic would give each output node a probability of the input belonging to a specific node/class.

1.2.5 Transposed convolutions

A transposed convolution is an operation that upsamples an input matrix to a specific shape. These operations are employed in several neural network architectures, such as in U-NETs. These operations can be understood as the ”reverse” of a convolutional operation with the caveat that only the shape of the input is reversed, not the values. Similarly to convolution operations, transposed convolutions also have kernels that are learnable parameters in neural networks. This process is depicted in *Figure 8.* By comparing *Figure 8* and *Figure 5* the conservation of shape, but not values, becomes clear.



135	69	154	207
104	79	187	230
288	189	217	124
260	265	229	205

1	0	1
0	1	0
1	0	1

135	0	135			
0	135	0			
135	0	135			

135	69	135	69		
0	135	69	0		
135	69	135	69		

135	69	289	276	154	207
104	214	360	463	394	230
423	293	873	707	601	331
364	632	969	996	540	435
288	449	770	542	422	124
260	265	489	470	229	205

Figure 8: A transposed convolution operation using a 3x3 kernel on an input tensor of size 4x4 with a stride of 1, and no padding leading to an upsampling of the image, and a corresponding output tensor of size 6x6. The value of each element in the input tensor serves as a weight for the kernel values. The kernel is overlapped with the output tensor and each value of the kernel is summed to the corresponding element in the output tensor. (A) The initial position of the kernel on the output tensor (B) The kernel moving one position to the right, due to the stride being 1. The overlapping values are summed. (C) The kernel iterates over the whole output tensor.

1.2.6 Multi Layered Neural Networks

Multilayered neural networks, often referred to as deep neural networks, consist of a sequential series of layers designed to process and transform input data into meaningful outputs. Each layer in these networks is composed of multiple neurons, and the structure of the layers plays a crucial role in how the network learns and generalizes from data. These layers can vary in type and function, depending on the specific architecture of the neural network.

Most important for this thesis are CNNs, composed of mostly convolutional layers. Unlike fully connected layers, convolutional layers employ a sparse connectivity pattern, where neurons are connected only to a small region of the input at a time. This design enables the network to detect local features such as edges, textures, and patterns in the data. These localized features can then be combined in deeper layers to identify more complex structures like objects or faces. This type of network excels at efficiently processing high-dimensional data, such as images, with relatively few parameters, reducing the risk of overfitting and improving computational efficiency.

Connecting the components mentioned until now, a network similar to the one displayed in *Figure 9* would be obtained.

1.2.7 Losses & Backpropagation

After the forward pass, the neural network needs to evaluate its performance to learn and successfully complete the required task. This evaluation involves measuring the accuracy or ”fidelity” of the output tensor with respect to the expected result, which is typically derived from the input data or other metrics set by the user. Quantification of the ”degree of fidelity” is done through a loss function. The loss function is a hyperparameter set by the user. There are numerous different loss functions, each of them is indicated for a specific context and will shape the model to generate wildly different outputs. For example, in multi-class classification , the most common loss function used for this problem is the cross-entropy loss function, also called log loss.

The cross-entropy loss function is given by:

$$\ell(x, y) = L = \begin{pmatrix} l_1 \\ \vdots \\ l_N \end{pmatrix}, \quad l_n = -w_{y_n} \log \left(\frac{\exp(x_{n,y_n})}{\sum_{c=1}^{C} \exp(x_{n,c})} \right), \tag{2}$$

where:

- x, y x is the input and y is the prediction,
- w is the weight,
- C is the number of classes,
- N is the size of the batch

This loss seeks to measure the difference between two probability distributions for a given set of events or values.

The loss function outputs a value that is passed through the network during the backwards pass, based on which learnable parameters, such as the kernels and weights, can be adjusted. The updating of these learnable parameters is usually done with an algorithm called stochastic gradient descent [28]. This optimization algorithm works in conjunction with backpropagation to update the learnable parameters in such a way that the loss is minimized. To this end, the negative gradient of the loss function is calculated. Then, each learnable parameter is updated in the direction of steepest descent by an arbitrary amount defined by the user through a specific hyperparameter called the ”learning rate”.

Thus, the equation for each weight in the network becomes:

$$w = w_0 - \alpha \frac{\partial L}{\partial w_0}, \tag{3}$$

where:

- w is the new weight of a parameter,
- w_0 is the old weight of a parameter,
- α is the learning rate set by the user,
- $\frac{\partial L}{\partial w}$ is the partial derivative of the loss function with respect to the weight.

To calculate the gradient at each layer in the network, the training process makes use of the backpropagation algorithm. The algorithm is based on the chain rule of calculus, which allows the gradient of the loss function to be propagated backward through the network. The chain rule is particularly useful in deep learning because it enables the calculation of the gradient of a composed function by multiplying the gradients of its constituent functions.

Mathematically, if the loss L depends on a parameter w through an intermediate variable z, and z itself depends on w, the chain rule states that:

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial w} \tag{4}$$

In a neural network, this concept is applied layer by layer, where the gradient of the loss function with respect to the weights of each layer is computed by multiplying the gradients from the subsequent layers. This allows for the updating of weights during backpropagation, ensuring that the loss is minimized as the network learns.

Equation 3 represents the most basic optimization algorithm. However, there are several different optimizers such as Adaptive Moment Estimation (ADAM). More advanced optimizers, like ADAM, make use of different, more complex techniques not covered in this thesis. For further details on ADAM and other optimizers, see [29]

After this step of backpropagation, the backwards pass, seen in *Figure 9*, through the network is complete [28]. The network will keep running this loop and training for as many batches as the user has set. The epoch and batch are also hyperparameters set by the user. The batch sets how many complete passes (forwards & backwards passes) are made by the network over the data or parts of it. The epoch can differ, but it generally refers to a complete pass over the whole data given to the model.

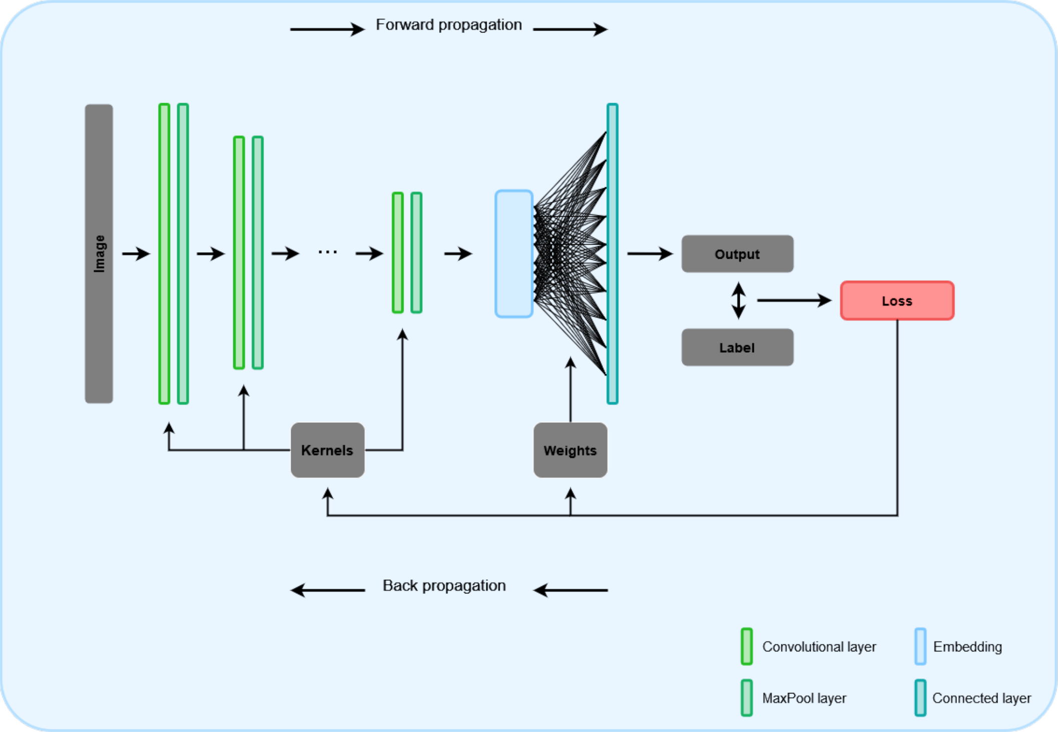


Figure 9: The basic architecture of a CNN. The information of the image given as input to the network flows through several convolutional layers that reduce the size of the image and increase the number of channels. After the convolutional layers, the resulting embedding is passed through a fully connected layer. Then the output is compared to a label through a loss function and a loss is computed. This loss is used to update the weights of the fully connected layer, and the kernels of the convolutional layers during back propagation

Finally, the network can be tested using a specific performance metric. Each test is designed by the user and tailored specifically to the task at hand. For a multi-class classification task, the performance metric might be the amount of correct classifications done by the network. Should the accuracy be satisfactory to the user, the model is kept. In the case where the performance does not attain a desired level, the network and the training procedure can be adjusted by modifying their hyperparameters, and retrained.

1.2.8 Datasets in neural network training

For the training of neural networks, 3 separate datasets are used. These datasets are derived subsets of the original, larger dataset intended for model training [30].

- **The training dataset.** A set of samples that is used during training to fit the model.
- **The validation dataset.** A set of samples used to tune the hyperparameters of a model and to perform a generalization check.
- **The test dataset.** A set of samples used to evaluate the final model.

An important factor in the training of effective and performant networks is the dataset, or the data, that is fed to the network. The most important characteristics of a dataset are its size and its quality. The size of the dataset needs to be sufficiently large to train the model, and of sufficient quality to avoid introducing biases to the model that affect the output.

Medical imaging datasets are usually small and not curated for DL applications. They usually lack quality labels due to the nature of human labeling, have different data types, are not stratified across important characteristics (age, gender), have occurrences that might be over-represented amongst the studied population, and are generally biased. [31]

This thesis focused primarily on a specific type of bias commonly found in medical imaging datasets, namely representation bias. These refer to a case when one or several parts of the input space are underrepresented. This can arise due to several reasons, such as [32]:

- **The sampling methods only reach a portion of the population.**
- **The population of interest has changed or is distinct from the population used during model training.** The population containing a certain disease is not sampled in accordance to known characteristics of the disease. For instance, hepatocellular carcinomas are situated in the right lobe of the liver with an occurrence of 75%, and the dataset used in training should represent this.[33]

1.3 Objectives

The objectives of this thesis are twofold. Firstly, it strives to analyse several facets of glioma distribution in MRI datasets. For each dataset, it seeks to analyse the distribution of tumors in the whole brain, and to determine which kind of registration, rigid or deformable, is most indicated for this type of analysis. Then it aims to also determine the tumor distribution in the cortical and subcortical, to gauge which structures are most affected in each dataset and how much this correlates with the distribution that is observed in nature. Finally, for this section, it endeavours to analyse how the connectome of the brain is affected in each dataset.

Secondly, this thesis compares two DL architectures, SNN and U-NET, and several different loss function to determine which of these architecture-loss combinations is best in the task of recognizing duplicate MRIs even when these have been augmented.

2 Materials & Methods and Results

2.1 Materials & Methods - Glioma distribution

2.1.1 URT & Data Collection

The Unified Retrieval Tool (URT) [34] is a tool, developed in house at the LCSB by the IAI team, for automatically downloading datasets from diverse sources and converting them to the standardized Brain Imaging Data Structure (BIDS) format [1]. This tool leverages the APIs of large open-source MRI dataset repositories to seamlessly download whole MRI dataset collections, such as Synapse and The Cancer Imaging Archive (TCIA), onto a computational environment, such as the High-performance Computer (HPC), with one simple command. Through this tool seven datasets were downloaded onto the HPC from two different repositories. In addition to the streamlining of the downloading process, the tool also converts the format of the datasets into the BIDS format wherever possible. The BIDS is a convention for the standardization of datasets containing medical images obtained from neuroimaging experiments. It was created to address the lack of standardization in the medical imaging community, and facilitate the reproducibility of experiments. The necessary commands used to download these datasets can be found in the IAI github repository.

The selection of the datasets was based on several characteristics. To qualify for analysis and inclusion in the experiment, the vast majority of the patients in the dataset had to contain at least a T1 image with its accompanying tumor segmentation. Since, most protocols use T1 images for patient registration to a standard space [35][36], and a tumor segmentation is required for the subsequent analysis. Alternatively, a T1, T1-contrast, T2, FLAIR, combination and no tumor segmentation is also acceptable, because a segmentation can be generated if four MRI modalities are provided. Furthermore, the size of the voxels, and general resolution of the images in the dataset was considered, with a low resolution and large voxels being a negative trait that would disqualify a dataset from being used. In general, a good dataset would contain a T1 image for each patient with its accompanying tumor segmentation, a voxel size of 1x1x1, and a resolution high resolution containing at least 80 slices. These factors also enabled the full utilization of the tools provided by Cancer Imaging Phenomics Toolkit (CaPTk), described later in this chapter. Since most datasets differ in their image acquisition protocol, this broad criteria allowed us to maintain a larger quantity of datasets that would otherwise not be possible. Finally, seven datasets were chosen: BraTS-2021, BraTS-SSA, LGG, RHUH-GBM, UCSF-PDGM, QIN, and Burdenko-GBM. The final list of datasets used for this part can be found in *Table 1*

Dataset	Modalities	# Patients	Seg.	Voxel Size	Resolution
BraTS-2021	T1, T1-contrast, T2, FLAIR	1251	✓	1x1x1	240x240x155
BraTS-SSA	T1, T1-contrast, T2, FLAIR	60	✓	1x1x1	240x240x155
Burdenko-GBM	T1, T1-contrast, T2, FLAIR	139	✗	variable	variable
LGG	T1, T2	150	✓	variable	variable
QIN	T1	48	✓	variable	variable
RHUH-GBM	T1, T1-contrast, T2, FLAIR	39	✓	1x1x1	240x240x155
UCSF-PDGM	T1, T1-contrast, T2, FLAIR	495	✓	1x1x1	240x240x155

Table 1: Summary of selected datasets, their modalities, number of patients, and tumor segmentation (Seg.) availability. For each dataset, each patient had only one tumor.

2.1.2 CaPTk & Registration

Registration of a brain MRI images to a standard or normal space was necessary in this thesis, because each patient's MRI is typically acquired in a unique anatomical space, influenced by factors an individuals anatomical differences. To ensure meaningful comparisons across different patients, it's necessary to align or "register" these images to a common reference space. This helps prevent skewing of the spatial distribution analyses by removing spatial discrepancies, allowing for consistent interpretation across the dataset's population.

CaPTk is a software platform for analysis of radiographic cancer images [37]. The platform integrates several tools for performing various aspects of medical image analysis, focusing mainly on brain, lung, and breast cancer imagery. Several tools provided by the CaPTk platform were used for this experiment, namely;

- **BraTS Pre-processing Pipeline.** The BraTS Pre-processing Pipeline is a pipeline designed around the popular BraTS dataset. It requires four structural MRI images (T1, T1CE, T2, FLAIR), in NIfTI format. The whole pipeline has four steps. Firstly, it reorients the image to RAI. Then it registers the images to the SRI24 atlas, by first doing an MRI-machine-introduced-bias correction, then a rigid registration of T1, T2, FLAIR to T1CE, and rigid registration of T1CE to SRI24. Finally, it applies the transformation to the reoriented images. Additionally, the BraTS Pre-processing Pipeline also outputs a skull and tumor segmentation created by the DeepMedic model.

- **DeepMedic.** DeepMedic is a DL model specialized in brain, and tumor segmentation [38]. It requires 4 basic MRI modalities (T1, T1-Gd, T2 and T2-FLAIR) for a subject which are co-registered. Using DeepMedic by itself is different from using it through the CaPTk pipeline, and yields different results.

- **Greedy.** Greedy is a tool for fast medical image registration [39]. It is based on Advanced Normalization Tool (ANTS) [40], which was also considered as a registration tool for use in this thesis. However, ANTS proved to be too computationally slow, and, since it was originally developed for use in Matlab, the Python wrapper was inadequate. These reasons made it difficult to work with ANTS at scale. In contrast, Greedy, was developed to be a fast CPU-based deformable image registration tool that could be used in applications where many images have to be registered in parallel.

Although, CaPTk has many features and tools incorporated in an easy to navigate environment, it has the disadvantage of not having been intended for use in pipelines that require many images to be run sequentially. For this reason, a CaPTk wrapper was written that can be used to easily utilize CaPTk's tools in an HPC environment, and manage large amount of images. Hence, for this thesis, each registration was done using Greedy through CaPTk. Greedy requires the following, user determined parameters to run:

- **Multi-resolution schedule (-n):** the number of iteration that Greedy's optimization algorithm is allowed to run at every resolution. For example: 100x40x20 does 100 iterations at the lowest resolution, 40 iterations at intermediate resolution, and 20 iterations at high resolution. This step tries to optimize the similarity metric.

- **Similarity Metric (-m):** used to calculate the similarity between two objects. They are used to quantify to what extent a transform is mapping the reference object on top of the target object. Greedy allows for three similarity metrics: Normalized mutual information (NMI), Sum of squared differences (SSD), and Normalized cross-correlation (NCC). Should NCC be used, then the user also has to define the *radius* metric, which defines the neighborhood of each voxel.

- **Initial transform specification for affine/rigid mode (-ia):** defines the initialization of a rigid/affine registration. Can be a given matrix or an identity matrix. Using the identity matrix will initialize the image alignment based on image headers. Using a given matrix will align the image centers.

After testing the registration of MRI images using several configurations, each registration in this thesis was done using the parameters listed below. The selection of this configuration was informed by literature where possible, or determined experimentally if not. For instance, the similarity metric was chosen to be NMI since it gives the best registration results [41], and the multi-resolution schedule, and initial transformation specification were determined through testing.

- **Multi-resolution schedule (-n):** 100x50x10.
- **Similarity Metric (-m):** NMI.
- **Initial transform specification for affine/rigid mode (-ia):** alignment based on image centers. -ia-image-centers.

Each dataset was registered rigidly and deformably to the two different MNI152 templates, MNI152 2009a NLIN asymmetric T1, and MNI152 2009c NLIN asymmetric T1. The MNI152 2009a template was used to calculate the spatial distribution of the tumors in the dataset, and the MNI152 2009c was used for the connectome analysis. These two registration steps were required to accommodate the connectomic atlas [42], which had been created around the MNI152 2009c template. In addition to this, the inverse warp generated by the deformable registration was used to register the Harvard-Oxford cortical and subcortical structural templates to each patient. These registrations were done to determine which structures are attained by tumors in each patient. An overview of all the registrations done for this section can be seen in *Figure 10*

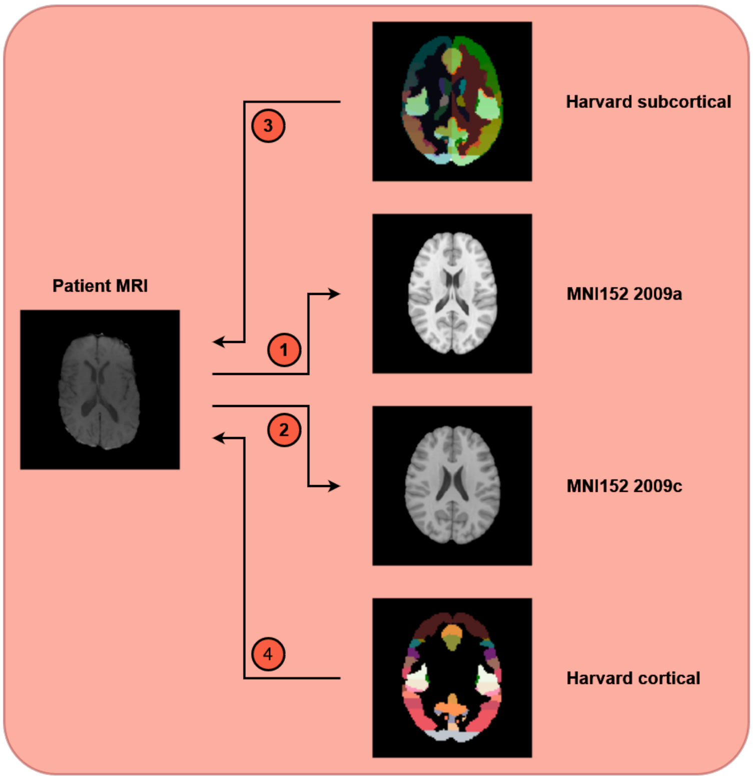


Figure 10: Overview of the registrations performed in this section. The arrows indicate the image that was registered to a target. The names of the templates were shortened for brevity. Path 1: Registration of the patient MRI to the MNI152 2009a NLIN asymmetric T1 template. Path 2: Registration of the patient MRI to the MNI152 2009a NLIN asymmetric T1 template. Path 3: Registration of the Harvard-Oxford subcortical template to the patient space. Path 4: Registration of the Harvard-Oxford cortical template to the patient space.

For the datasets that have no tumor segmentations, such as the Burdenko-GBM dataset, segmentations had to be generated. The tumor segmentations were generated by using the BraTS Pre-processing Pipeline. This did not require any specific arguments to be given, except for the images mentioned the tool requires that were mentioned previously. The BraTS Pre-processing Pipeline also registers the images to the SRI24 brain atlas, however, since we set on working with the MNI152 2009a NLIN asymmetric T1, a subsequent registration step to this template had to be done for the tumor segmentations and T1 images of these datasets. The tumor segmentation was also done outside the BraTS Pipeline, through DeepMedic, and the best registration was selected.

2.1.3 Calculation of tumor distribution in the MNI space

For the tumor distribution in the MNI space, each image in each dataset was registered to the MNI152 2009a NLIN asymmetric T1 template. This corresponds to path 1 in *Figure 10.* Each image was registered twice, once deformably and once rigidly.

Commonly, tumor segmentations consist of three labels, the edema, the enhancing region, and the necrotic core. An example can be seen in *Figure A.1* in the appendix. The edema was ignored for this analysis, leaving only the enhancing region and the necrotic core. This was done, because the edema is an accumulation of fluid and the analysis targeted only tumor tissue. Moreover, since, the enhancing region is usually very small, the necrotic core and the enhancing region were considered as one label, referred to from now on as the "core". This created essentially a binary mask, for which the value was set to one. This process was repeated for every patient in each dataset. For each dataset, the tumor core binary masks were added together through simple addition, and the divided by the total number of patients in that dataset. Creating a 3-dimensional spatial distribution volume for each dataset. Similar methods have been applied before [43]

To create a 2-dimensional probability distribution map for each dataset, the same method of binary tumor core addition was performed. Then the 3-dimensional volume was collapsed into 2 dimensions by adding together the values in each layer, from top to bottom. Finally, each value of the 2-dimensional array was divided by the total number of voxels in the 3-dimensional volume. Shown in *Equation 5.* Creating a 2-dimensional probability distribution map.

$$P(x,y) = \frac{1}{N}\sum_{z=1}^{Z} V(x,y,z) \tag{5}$$

where:

- $P(x,y)$ is the value at position (x,y) in the 2-dimensional probability distribution map,
- $V(x,y,z)$ is the value at position (x,y,z) in the 3-dimensional volume,
- Z is the total number of layers in the 3-dimensional volume (155),
- $N$is the total number of voxels in the 3-dimensional volume ($155\times240\times240$)

2.1.4 Calculation of regional tumor distribution in the native space

For the regional tumor distribution in the native space, the Harvard-Oxford subcortical & cortical templates were registered to the patient's MRI space. This was done for several reasons. During registration, especially deformable registration, of a patient's image it is inevitable that the tumor tissue will be deformed and moved. To calculate the tumor distribution across the whole brain, this isn't very important because the tumor can't be registered to the outside of the brain. However, when considering smaller structures, such as brain regions, small induced deformations of the tumor can change the region in which the tumor is located in. To exemplify, a tumor in the primary somatosensory cortex can be deformed in a way that would locate part of its tissue in the primary motor cortex, falsifying the analysis. Registering the template to the patient's brain, path 3 & path 4 in *Figure 10* hopes to lessen these errors.

The calculation of the percentage of the tumor contained within each region was done similarly to the method described previously. Firstly, for each patient in a dataset, the edema was ignored, and only the tumor core retained for analysis. Then, the voxels of the tumor contained within a specific region were added and divided by the total number of voxels of the region, *Equation 6*. This yielded the percentage of tumor tissue befalling each region. Finally, these individual percentages were used to calculate the statistics for the whole dataset, *Equation 7*.

$$\text{Tumor Fraction in Region} = \frac{\sum \text{Voxels}_{\text{Tumor in Region}}}{\text{Total Voxels in Region}} \tag{6}$$

$$\text{Dataset-Wide Tumor Fraction per region} = \frac{\sum_{i=1}^{N} \text{Tumor Fraction in region}_i}{N}, \tag{7}$$

where:

- N is the number of patients in the dataset

2.1.5 Calculation of connectome attainment

As previously mentioned, gliomas are intimately entangled with the brain's connectome. Additionally, studies suggest that glioma lesions located within regions that contain a higher density of white matter fibers indicate a lower survival expectancy for the patient. Furthermore, recent paper have provided a populational, probabilistic multi-scale atlas of the connectome based on Diffusion Weighted Imaging (DWI) and tractography data, that allow for performance of several kinds of brain connectivity analysis even in absence of DWI and tractography data [42]. In this atlas, the value of each voxel corresponded to the average number of white matter tracts passing through it. In the context of the datasets used in this thesis, that did not have DWI modalities, this allowed for analysis of the impact of glioma on the connectome in each of the acquired datasets.

To calculate the connectome attainment, the tumor segmentation of each patient in a dataset was registered to the MNI 2009c NLIM asymmetric T1. Corresponding to path 2 in *Figure 10*.

Here too, the edema was not considered, and the enhancing region and necrotic core were treated as a single segmentation. The binary mask of the tumor segmentation was overlaid with the multi-scale atlas of the connectome. Then only the overlap of the connectome and the tumor segmentation were considered. Each voxel in this overlap was added. The total was divided by the total number of voxels. Yielding the average number of white matter tracts attained by the tumor for each patient. The severity of the impact of the tumors on the connectome was considered low, if the number of white matter tracts attained by the tumor was less than 100. It was considered medium if the number of white matter tracts attained was over 100, but under 250, and high if it was over 250.

2.2 Results - Glioma distribution

Deformable and rigid registration of the patient's T1 MRIs to the MNI152 2009a NLIN asymmetric T1 template, path 1 in *Figure 10*, and summation of all the resulting registered patient images across a dataset as described in the chapter *Calculation of tumor distribution in the MNI space*, yielded the results seen in *Figure 11.* To ease visualization, the resulting 3D spatial tumor distribution images were simplified into three 2D slices. These slices correspond to the axial, sagittal, and coronal views, centered on the point of highest tumor incidence probability. The methods used for registration, deformable and rigid, were compared by overlapping both images, *Figure 11 (C, F).* This comparison shows strong differences between rigid and deformable registrations, impacting even the points of highest tumor incidence probability. This can be seen in *Figure 11 (Top: A, D* and more starkly in *Figure 11 (Bottom: B, E)*, where the point of highest tumor probability shifts completely between rigid and deformable registration. Moreover, after rigid registration, the tumor distribution overlaps with several brain structures in the MNI152 2009a template. In contrast, the deformable registration minimizes these overlaps, and the distribution remain in anatomically sensible locations. Highlighted structures in *Figure 11* such as the ventricles and the brain sulci are not overlapped after the deformable registration. Further comparisons can be found in *Figure A.24* in the appendix.

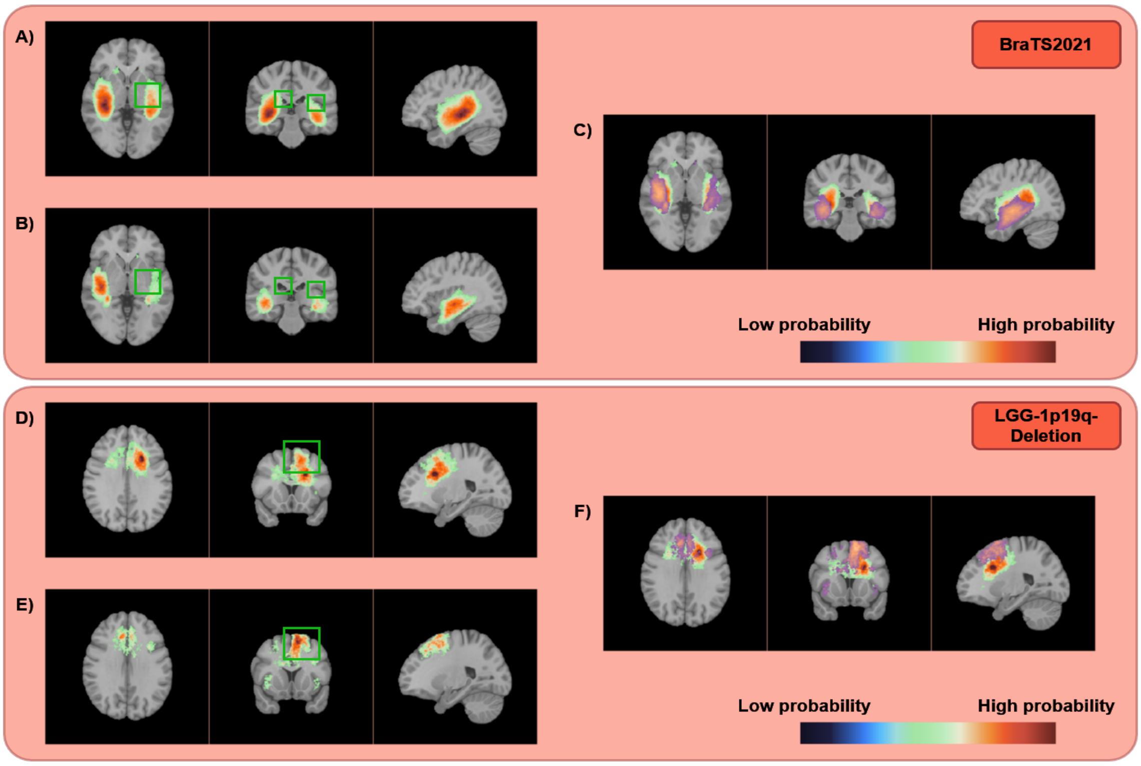


Figure 11: Results of the tumor distribution in MNI space. (Top) Spatial distribution of tumors in the BraTS-2021 dataset. (A) Axial, coronal, sagittal views of the spatial tumor distribution after rigid registration of the MRI images to the MNI152 2009a template. (B) Axial, coronal, sagittal views of the spatial tumor distribution after deformable registration of the MRI images to the MNI152 2009a template. (C) Axial, coronal, sagittal views of the overlap between the spatial distributions after rigid and deformable registration. The results from the rigid registration kept the jet color scheme, while the results from the deformable registration were changed to the copper color scheme to enhance contrast. (Bottom) Spatial distribution of tumors in the LGG dataset. (D), (E), and (F) are the same as their counterparts in (Top). The green boxes in both images show examples of brain structures that are overlapped after rigid registration and not after deformable registration

More generally, a shift in the spatial distribution from one location to another depending on the registration method was also perceived, *Figure 12*. Although, the tumor distribution did not change between brain hemispheres, within a hemisphere the shift in method from rigid to deformable made the distribution more pointed to a specific region of the brain. *Figure 12 (Top)* shows a blunted distribution peak obtained after rigid registration turn into a sharp peak after deformable registration. The remaining plots for each dataset can be found in the appendix (*Figure A.23*).

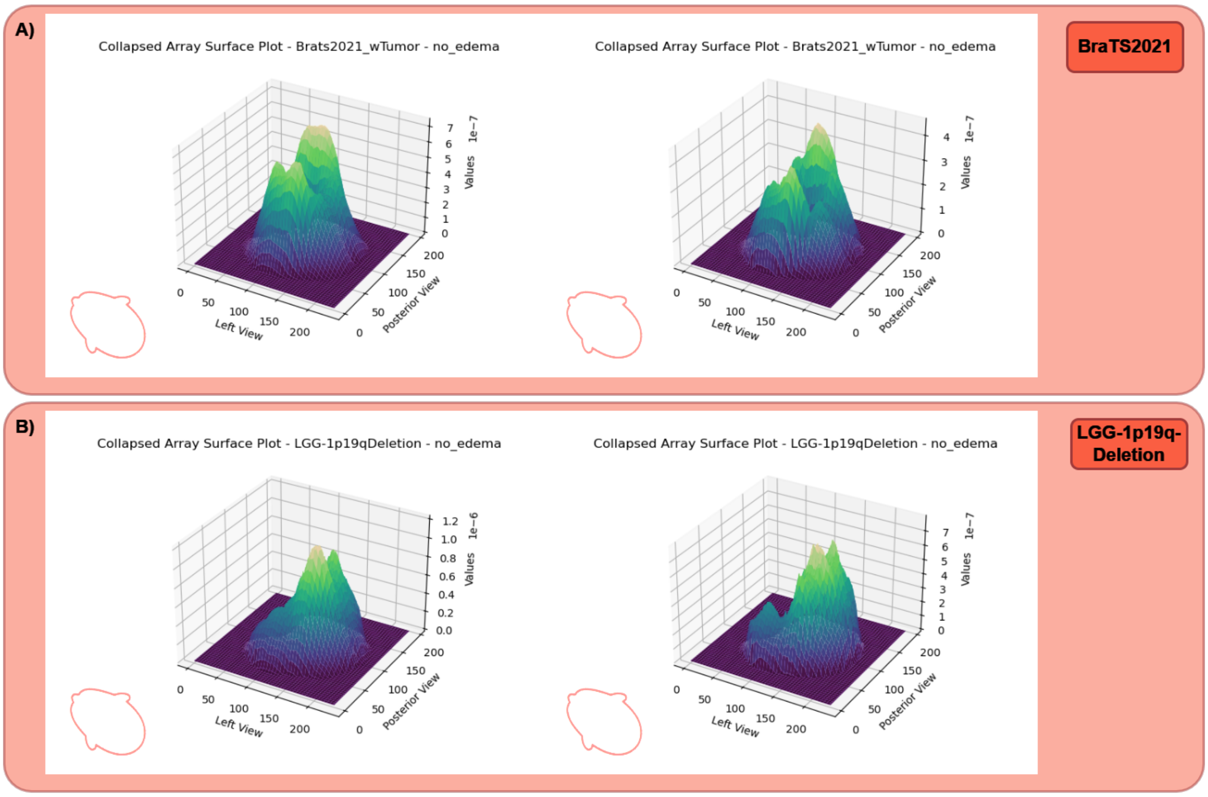


Figure 12: (Top) The probability distribution in 2D of the BraTS-2021 dataset. (Bottom) The probability distribution in 2D of the LGG dataset. (Left) Probability distribution after rigid registration to the MNI152 2009a. (Right column) Probability distribution after deformable registration to the MNI152 2009a. The "Posterior View" axis indicates the posterior side of the brain, and the opposing axis to this one is the anterior view of the brain. The "Left View" axis indicated the left side of the brain, and the opposing axis to this one is the right view of the brain.

Between the cortical region, deep subcortical structures, and subcortical white matter, the tumors preferentially attained the subcortical regions in most datasets, as is shown in *Figure 13* and *Table 2*. Interestingly, the LGG showed the largest contrast between deep subcortical structure attainment and attainment of other regions, with only 2.9% of tumor attaining deep subcortical structures. Furthermore, this trend remains independent of dataset size, with larger datasets showing the same preference. The highest proportion of deep subcortical attainment amongst all datasets was found in the BraTS-SSA dataset that showed a subcortical attainment of 17.5%. The contrast between cortical region and subcortical white matter attainment is less stark. Most datasets show a slight preference for either region. The starkest contrast is in the UCSF-PDGM dataset, where 29.71% of tumors attain the cortical region, and 47.48% of tumors attain the subcortical white matter. The detailed plots for each dataset can be found in the appendix (*Figure A.10*, *Figure A.11*, *Figure A.12*, *Figure A.13*, *Figure A.14*, *Figure A.15*).

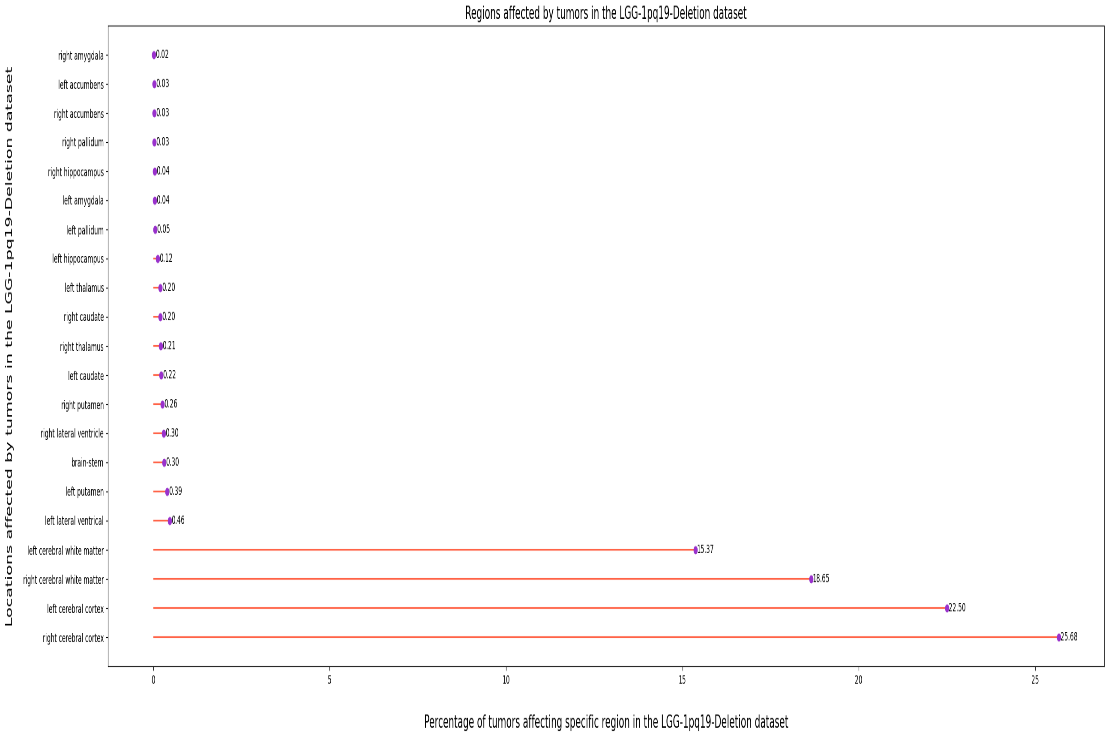


Figure 13: The subcortical regions, in contrast to cortical regions affected by tumors in the LGG dataset.

Dataset	**# Patients**	**Structure (%)**		
		Subcortical - DS	**Cortical**	**Subcortical - WM**
BraTS-2021	1251	7.77	43.17	46.45
BraTS-SSA	60	17.5	31.65	42.78
Burdenko-GBM	139	4.39	44.4	46.67
LGG	2.9	2.9	48.18	48.29
QIN	48	9.87	44.96	31.56
RHUH-GBM	39	6.6	42.74	50.09
UCSF-PDGM	495	6.81	29.71	47.48

Table 2: Summary of the regions affected by tumors in the other datasets. The percentage of attainment for each structure (Subcortical - DS, Cortical, and Subcortical - WM) is shown. DS refers to deep subcortical structures. WM refers to subcortical white matter. Missing percentages correspond to unassigned voxels. Summary corresponds to *Figure A.10*, *Figure A.11*, *Figure A.12*, *Figure A.13*, *Figure A.14*, *Figure A.15* in the appendix

The deep subcortical regions were differentially affected in each dataset. *Figure 14* shows a qualitative comparison of each deep subcortical structure, and its attainment, for each dataset.

Highlighted in this figure are the deep subcortical structures that show the greatest attainment discrepancy between all datasets. The brain stem shows the largest variability between all datasets, followed by the right hippocampus, and right caudate. Here again the BraTS-SSA dataset shows the largest attainment of the brain stem and of the right caudate amongst all datasets.

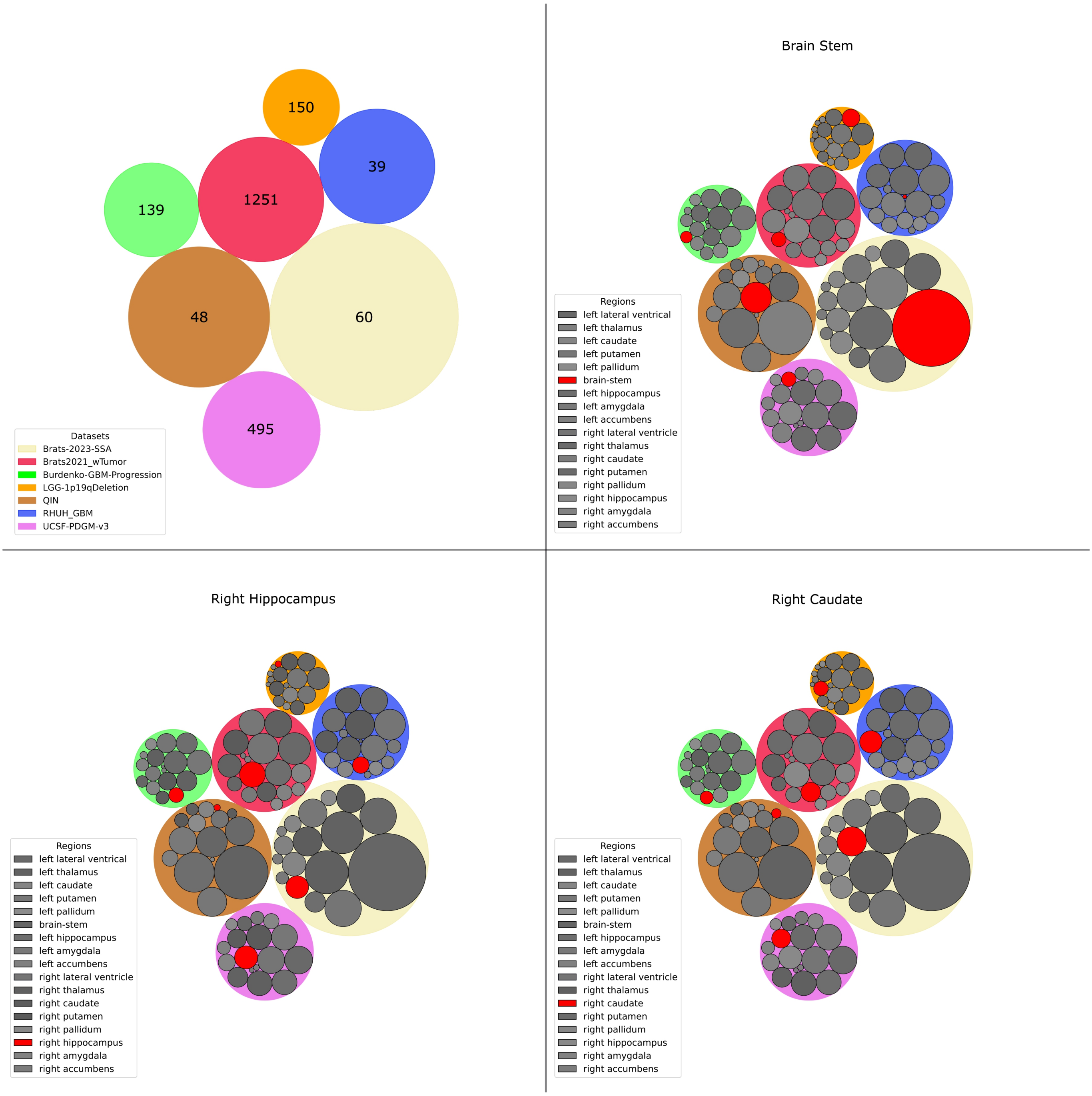


Figure 14: A qualitative comparison of attainment of deep subcortical structures by gliomas between the investigated datasets. The numbers inside of the large circles are the number of patients for that dataset. The size of the large circles corresponds to the proportion of subcortical attainment by tumors for that dataset. The size of the small circles corresponds to the proportion of attainment of that subcortical structure by tumors.

Similarly to the deep subcortical regions, each cortical region in each dataset was attained differently, as can be seen in *Figure 15*. However, unlike with the subcortical structures, no discernible patterns emerge for the cortical regions. *Table 3* gives a summary of all the most important cortical regions, and the percentage of tumors that attained them.

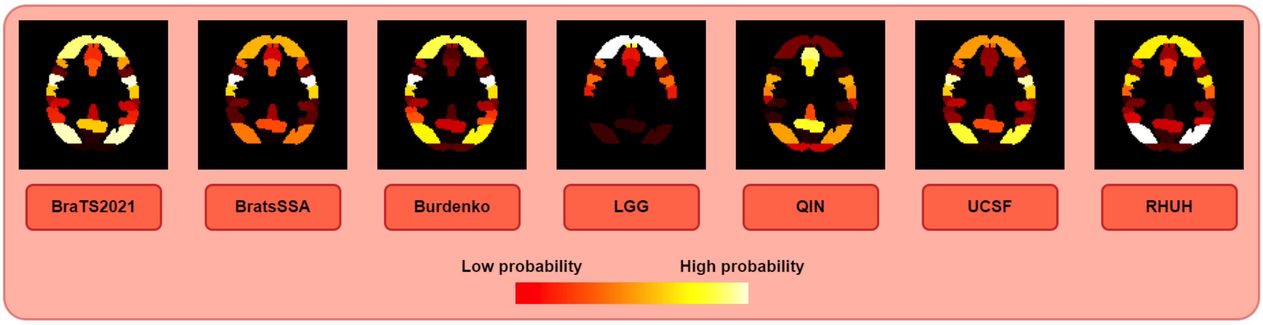


Figure 15: The probability of tumors impacting a specific cortical subregion in each dataset. Brighter colors indicate a higher probability of a region being impacted by tumors.

Dataset	**Region**		
	Temporal	**Occipital**	**Frontal**
BraTS-2021	12.32%	3.89%	7.68%
BraTS-SSA	7.88%	2.44%	3.43%
Burdenko-GBM	14.27%	4.88%	7.01%
LGG	23.95%	1.50%	3.13%
QIN	7.96%	3.92%	4.77%
RHUH-GBM	11.95%	6.53%	8.11%
UCSF-PDGM	9.67%	4.08%	8.28%

Table 3: Overview of the main cortical regions attained by tumors. Corresponding to *Figure A.16*, *Figure A.17*, *Figure A.18*, *Figure 17*, *Figure A.20*, *Figure A.21*, *Figure A.22* in the appendix.

The hemispheres in each dataset also showed differential attainment. Most of the datasets show a stark uneven distribution of tumors between both hemispheres. Four datasets show a preference for the left hemisphere, and three show a preference for the right hemisphere. The BraTS-SSA dataset is the most evenly distributed, having only a very light preference for the right hemisphere. Again, the size of the dataset seems to play no role in the preference for one hemisphere over the other. BraTS-2021 has the highest patient count, and the left side is also preferentially attained by tumors. However, it's noteworthy that the smaller datasets, such as RHUH-GBM, and QIN show the strongest disparity between both hemispheres.

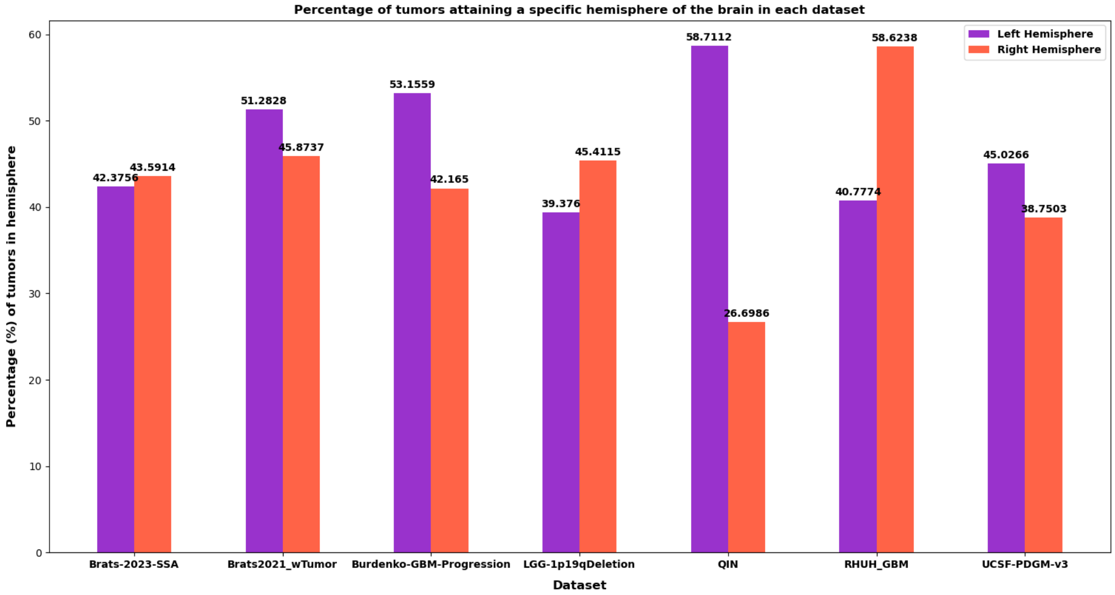


Figure 16: The percentage of tumors falling into a specific hemisphere of the brain. The most striking differences occurs in datasets with lower patient counts (nRHUH-GBM = 39 , nQIN = 48), but are not the rule (nBraTS-SSA = 60). Larger datasets tend to have a more even distribution, however a bias for one hemisphere over the other remains

The impact of the tumors on the connectome differed in severity for each dataset, as can be seen in *Figure 17* and *Table 4*. The greatest proportion of low severity attainment was found to be in the QIN and Burdenko-GBM datasets. Each having a low severity proportion of 48%. The greatest medium severity was found in the RHUH-GBM dataset with 74%. Finally, the greatest high severity proportion was found the LGG dataset with 12%. The size of the dataset does not seem to play a role in the proportion of any distinct severity. The full plots belonging to the other datasets can be found in the appendix (*Figure A.4*, *Figure A.5*, *Figure A.6*, *Figure A.7*, *Figure A.8*, *Figure A.9*).

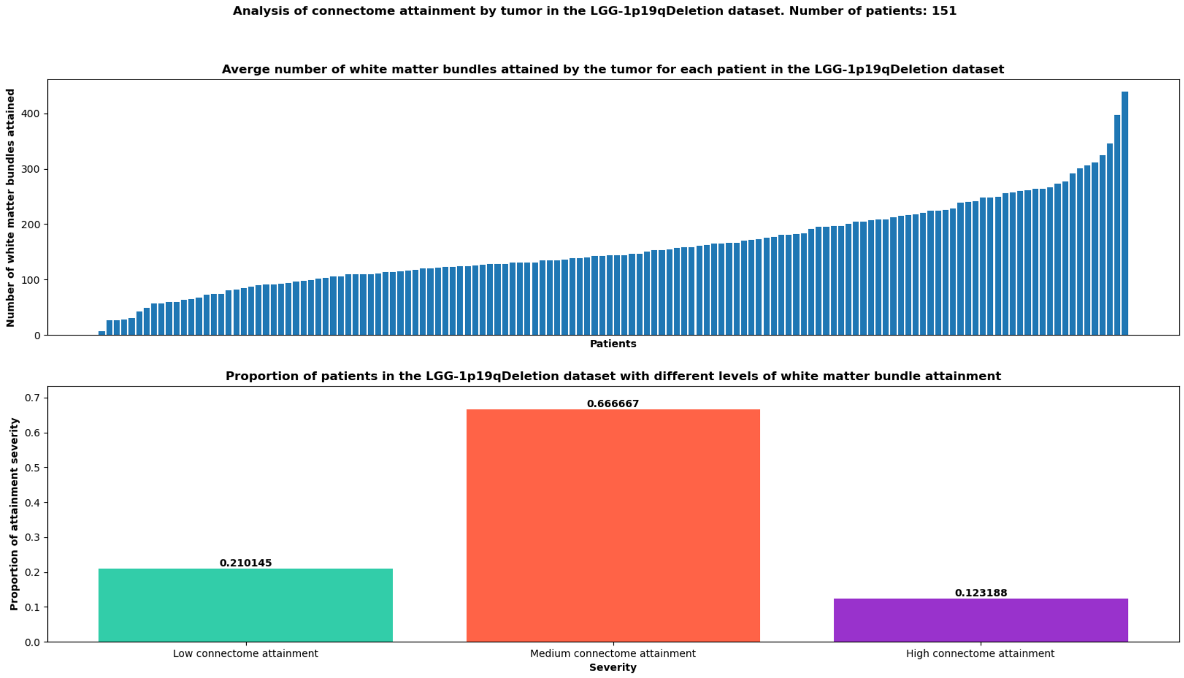


Figure 17: (Top) The number of white matter tracts attained by the tumor in each patient of the LGG dataset. (Bottom) The percentage of low, medium, and high severity of connectome attainment in the dataset. Low severity = white matter tracts attained by the tumor $<$ 100. Medium severity = 100 $<$ white matter tracts attained by the tumor $<$ 250. High severity = white matter tracts attained by the tumor $>$ 250.

Dataset	**# Patients**	**Severity in %**		
		Low	**Medium**	**High**
BraTS-2021	1251	42	54	4
BraTS-SSA	60	33	65	2
Burdenko-GBM	139	48	50	2
LGG	150	21	67	12
QIN	48	48	48	4
RHUH-GBM	39	21	74	5
UCSF-PDGM	495	42	54	4

Table 4: Summary of the severity of connectome attainment in the other datasets. Corresponding to *Figure A.4*, *Figure A.5*, *Figure A.6*, *Figure 17*, *Figure A.7*, *Figure A.8*, *Figure A.9* in the appendix

To summarize, this section found significant differences between the two methods of registration, particularly in how they affected tumor distribution in the brain. Deformable registration provided more anatomically accurate results, with less overlap between tumors and key brain structures compared to rigid registration. Across datasets, tumors predominantly affected subcortical regions, with varying degrees of impact on different brain structures. This section also examined how tumors differentially impacted the brain's hemispheres, cortical regions, and the connectome, finding variations in tumor distribution and severity across different datasets.

2.3 Materials & Methods - Duplicate Recognition

This section focuses on addressing the second objective of this thesis - Proposing a model capable of identifying duplicate MRIs, and investigating which losses work best for this task. This section covers both networks that were tested for the final model, the losses used, the image augmentations the model was trained to handle, and the performance of the model with the various losses.

2.3.1 Siamese Neural Network

The SNN is a special training procedure and loss function used in DL [44], depicted in *Figure 18*, that was originally developed for one-shot classification, meaning that a model is expected to classify an instance of a category based on only one example of that category. Usually, models that are trained for a classification task require numerous examples of each category to perform well. This particular characteristic of SNNs to identify the category of an instance with little examples makes it ideal for tasks such as airport baggage security checks, facial recognition, or for tasks wherein examples are scarce, such as medical images [45][44].

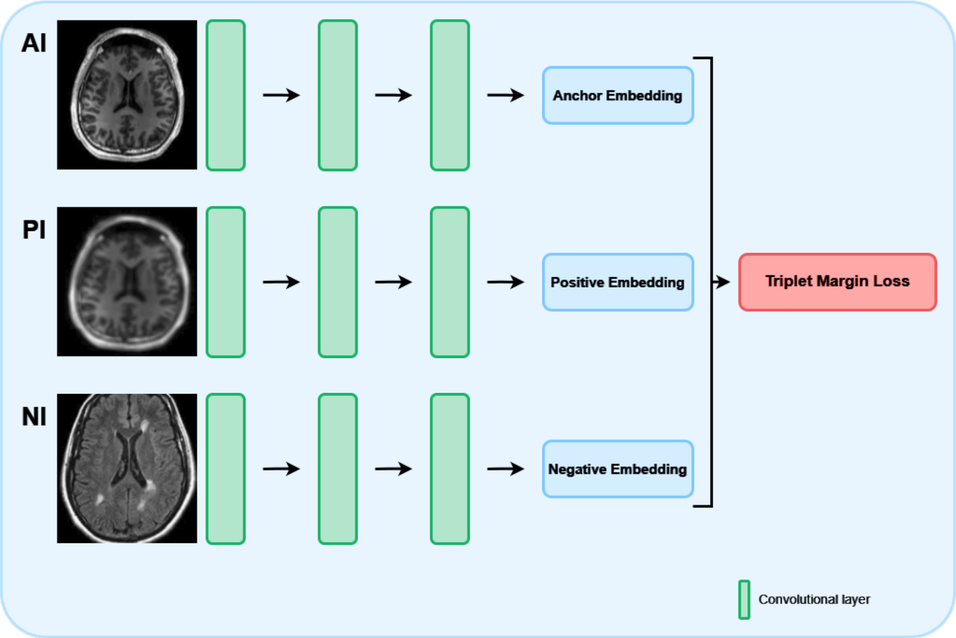


Figure 18: The basic training procedure and loss function of a SNN. The anchor image (AI), positive image (PI), and negative image (NI), here depicting MRI images, are passed through a convolutional network. The resulting embeddings from each image are used to calculate the triplet margin loss. This loss is subsequently back propagated through the network to update the weights of the network.

The architecture of the SNN used in this thesis was composed of 6 convolutional layers, extracting a total of 1024 features. The final layer of the SNN is a fully connected layer that yields an embedding vector of size (1, 1024), a high level abstraction of the original input. To obtain an embedding vector for each image, the original image (referred to as "anchor"), the augmented image of the original image (referred to as "positive"), and the image that is unrelated to the original image (referred to as "negative"), each image has to pass through the SNN. Finally the embedding vectors are used to calculate a loss using the triplet margin loss function, as is depicted in *Figure 18* [44].

Critical for the functioning of SNNs is a contrastive loss, also called embedding loss in this thesis, such as the triplet margin loss seen in *Equation 8*.

$$L(a, p, n) = \max\{d(a_i, p_i) - d(a_i, n_i) + \text{margin}, 0\}, \tag{8}$$

where:

- $d(x_i, y_i) = \|x_i - y_i\|_p$
- p is the norm degree for pairwise distance

The triplet margin loss maximizes the distance between the anchor embedding (a) and the negative embedding (n), and minimizes the distance between the anchor embedding and the positive embedding (p). The margin refers to a margin between the negative and positive pairs. Once back propagated, the triplet margin loss will allow the model to learn to distinguish between the images that are given to it. [46]

2.3.2 U-NET

The U-NET is a deep convolutional network originally created for cell segmentation [47]. However, this type of architecture also excels in medical imaging segmentation, where images are mapped to segmentation masks. Additionally, this architecture has proven to be very successful in all sorts of different tasks, such as super-resolution [48], where a low resolution image is upscaled to high resolution, or in diffusion models [49], where Gaussian noise, or noise in general, is turned into images.

For each of these tasks, both the input and the output are images. If we take the example of image segmentation, by having hand-labeled segmentation masks as ground truths, we can compare the output of the U-NET to the true labels and compute a loss based on a loss function, often a cross-entropy loss. Through backpropagation of this loss through the network, the U-NET will learn to create more accurate segmentation maps after each iteration.

The U-NET architecture, depicted in *Figure 19*, has two main components, the encoder going down the network, and the encoder going up the network. This gives the network its characteristic name; U-NET. The down-path, also called the contracting path, is responsible for extracting the features from the image, and the up-path, also called the expansive path, is responsible for restoring the spatial dimensions of the image and outputting a final image [47]. Both paths are symmetrical and contain multiple layers.

The encoder is made of sequential, and repeated layers. Each layer is composed of two 3x3 convolution operations followed by a ReLU activation function. Connecting each layer is a 2x2 pooling operation that works to downsample the spatial dimensions of the image. Then the features are doubled by the next layer, and so on until a user defined number of features is attained.

The decoder can be understood as the inverse of the encoder. Similarly, it is composed of sequential, and repeated layers, made up of a 3x3 convolution operation followed by a ReLU activation function. In contrast to the encoder, in the decoder the operation between the layers is not a downsampling operation, but an upsampling operation. The 2x2 transposed convolutional layer upsamples the spatial dimensions of the image, restoring the original image resolution, and the features are reduced by the next layer in the sequence.

There are two further components important in the U-NET. The skip connections and the bottleneck. The skip connections connect each layer in the encoder with its symmetrical opposite in the decoder. The skip connections take a copy of the features in the encoder path and concatenate it to the features of the decoding path. This prevents spatial information from being lost by injecting high resolution features directly to the decoder. Furthermore, the encoder captures more localized information, while the decoder has more semantic information and requires precise localization information. By merging these together through the skip connections, the network performs better in segmentation tasks.

The bottleneck is where the network switches from encoder to decoder. This component has special characteristics that are interesting for this thesis' application. It contains the final embedding vector generated by the downward path. This vector contains high-level, highly abstract information about the image.

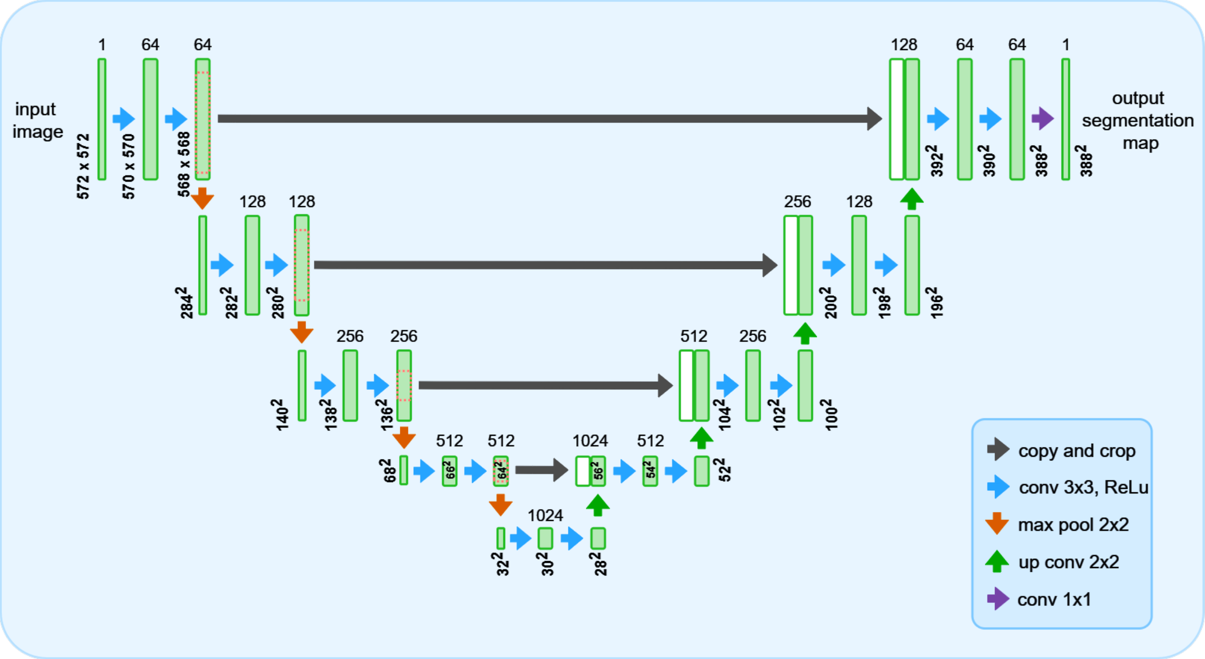


Figure 19: The classical U-NET architecture [47]. (Left) The encoder, downward path, composed of sequential convolutional layers and max pool operations, flows into the bottleneck, where a high-level, abstract embedding is contained. Then the network switches to (Right) the decoder, or upward path, composed of sequential convolutional layers and transposed convolutional operations for upsampling of the image. Between each layer of the encoder and decoder, a skip connection connects both and lets information flow from the encoder to the decoder.

A postulate of this thesis, rests in the idea that embedding vectors can represent images, and consequently be used to identify them. This was proposed in other bodies of work, but only using SNNs [50]. This thesis elected to use a U-NET to investigate if adding a loss based on the reconstructed image could make the embeddings of the U-NET more reliable than those of an SNN. The supposition for this was that the reconstruction losses could imbue more ”biological” information into the embedding, leading to a more performative model. To this end, the classical U-NET, sourced from the Medical Open Network for Artificial Intelligence (MONAI) framework, was slightly modified, and special losses were used, as is depicted in *Figure 20.* This modified U-NET, referred to as Modified U-NET (MU-NET) from here on out, takes three different images as inputs. It takes the image it is supposed to learn to recognize, also called the anchor image (AI). Secondly, it takes an augmented version of the AI, called a positive image (PI). The PI image is used to ensure that the network learns to recognize the AI, even when the AI is modified by an augmentation. The augmentations used were extensive and randomized so that no batch of images would have the same augmentations, forcing the model to generalize. Finally, it takes an image unrelated to the original image, called the negative image (NI). The NI ensures that the network learns to differentiate different images from the original image. The embedding vectors at the level of the bottleneck were extracted from the U-NET architecture for each of the three images and saved before passing them through the rest of the architecture, which is the modification done to the U-NET architecture in this thesis. At the decoder end of the network, three reconstructions of the three separate images were retained and used to calculate a loss

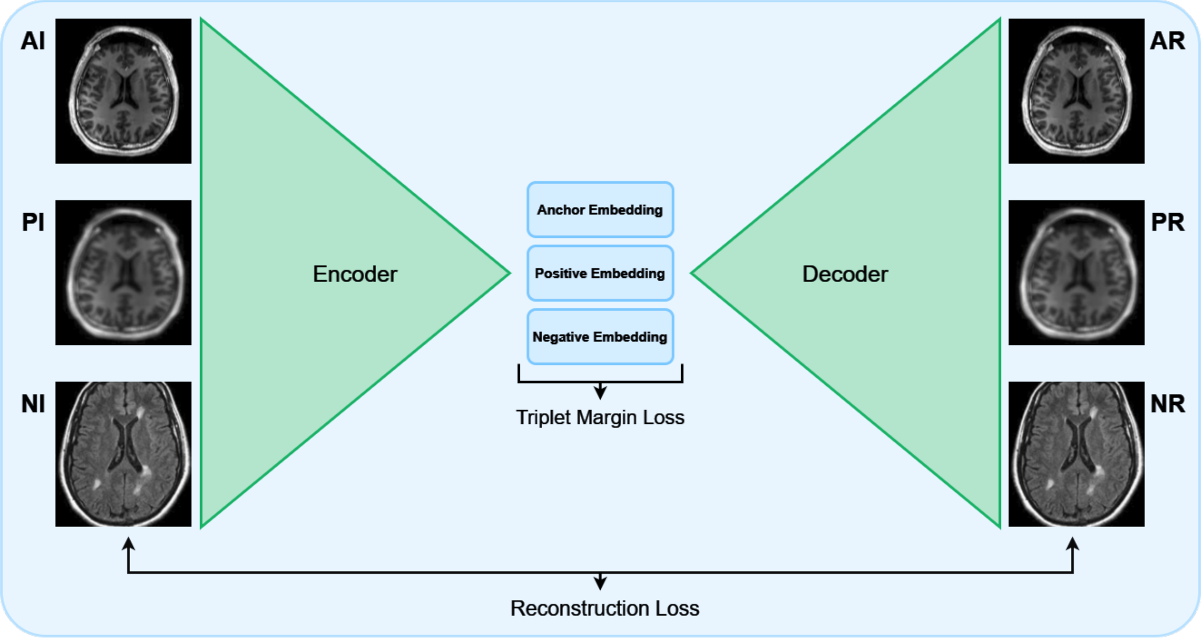


Figure 20: The classical U-NET architecture with an additional modification that allows the embeddings of the network to be extract and a loss to be calculated on them that can be used for backpropagation. (AI) Anchor Images, (PI) Positive Image, (NI) Negative Image. (AR) Anchor reconstruction, (PR) Positive reconstruction, (NR) Negative reconstruction.

In addition to the U-NET architecture from MONAI described above, a second U-NET network from MONAI, referred to as Basic U-NET (BU-NET) from here on out, was used for some of the experiments. The reason for using a second U-NET architecture was to try and determine the origin of an artifact that appeared on the reconstructed images of the MU-NET. The BU-NET has the exact same structure as the MU-NET. However, the transposed convolutions, with learnable kernels, on the upward path can be replaced by bi-linear interpolation layers.

2.3.3 Datasets

To train both of these architectures three datasets were used, each with 1251. One containing only T1 images, another containing only T2 images, and a third containing both T1 images and

T2 images. Each dataset was subdivided into a training set (70 %), a validation set (20 %), and a test set (10 %). For each network the training loss, and the validation loss were plotted using Weights & Biases, an online tool for plotting DL related metrics [51].

BraTS-2021 has images from previous BraTS challenges, such as Brain Tumor Segmentation 2020 (BraTS-2020). For this reason, BraTS-2020 was downloaded to test the performance of the model. There are 365 images from the BraTS-2020 dataset known to be included in the BraTS-2021 dataset. The whole size of the BraTS-2020 is 369 images, making it an ideal test dataset to test the performance of the model in a semi-real environment.

2.3.4 Augmentations

As mentioned previously, the aim of this section is to create a model that can reliably identify if an MRI is already contained within the dataset one wants to add it to. However, as mentioned in section 3.1.2 MRI images can undergo unintended augmentations as they are disseminated through several groups. To simulate this, several sets of augmentations were applied to the training data. Each set of augmentations refers to one experiment.

Extensive augmentations. A set of various augmentations, some of which are not expected to occur in MRIs, such as affine transformations, and blurring. These were added to increase the amount of available data.

- Bias Field errors
- Blurring
- Image noise
- Voxel anisotropy
- Image spikes
- Affine transformations
- Image flips

Compact with canonical augmentations (Compact + Canonical). A set of augmentations that are expected to be found after manipulation and re-dissemination of MRIs.

- MRI orientation switch to canonical. The canonical orientation refers to the RAS (Right, Anterior, Superior) orientation of an MRI.
- Image flips

- Voxel anisotropy
- Image noise

Compact without canonical augmentations (Compact).

- Image flips
- Voxel anisotropy
- Image noise

2.3.5 Experiments

In total three experiments were performed in this section.

Experiment 1: The first experiment was in identifying T1 images from augmented T1 images. Here the anchor image was a T1 image. The positive image was the same T1 image, but augmented with an augmentation set. The negative image was a different T1 image from the anchor image. This experiment was performed for each of the augmentation sets previously mentioned.

Experiment 2: The second experiment was in identifying T1 images from augmented T2 images. Here the anchor was a T1 image. The positive image was the T2 image from the same patient, but augmented with the compact augmentation set. The negative image was a different T2 image from the anchor image. This experiment was performed only on the compact augmentation set.

Experiment 3: The third experiment was in identifying T1 images from augmented T1 images or augmented T2 images. Here the anchor was a T1 image. The positive image was the same T1 image, or the T2 image from the same patient, but augmented with an augmentation set. This experiment was performed only on the compact augmentation set.

2.3.6 Performance Measure

The accuracy of each model was determined by measuring the number of images that were correctly identified after augmenting them, given in *Equation 9* [52]. The augmentations used for the augmented images were the same as those used during training. Furthermore, the accuracy was also measured when there were no augmentations of the images, and for different modalities. A detailed illustration of how the accuracy was measured can be found in *Figure 21.* To get a more robust measure of accuracy for any specific model, the model was run for five iterations during testing. Then the sum of the accuracy in each iteration was divided by the total number of testing iterations.

$$\text{Accuracy} = \frac{\sum_{i=1}^{\text{n}} \frac{\text{number of images correctly identified}}{\text{number of total images}}}{\text{total testing iterations}} \tag{9}$$

For each set of augmentations and losses, the accuracy of the model ended up being measured in two separate ways. $Accuray_{T1a}$, identifying T1 images from augmented T1 images. $Accuray_{T1}$, identifying T1 images from non-augmented T1 images. For the set of augmentations applied to T2 images the accuracy was calculated based on the T2 augmentation set. $Accuray_{T2a}$ identifying T1 images from augmented T2 images. $Accuray_{T2}$ identifying T1 images from non-augmented T2 images.

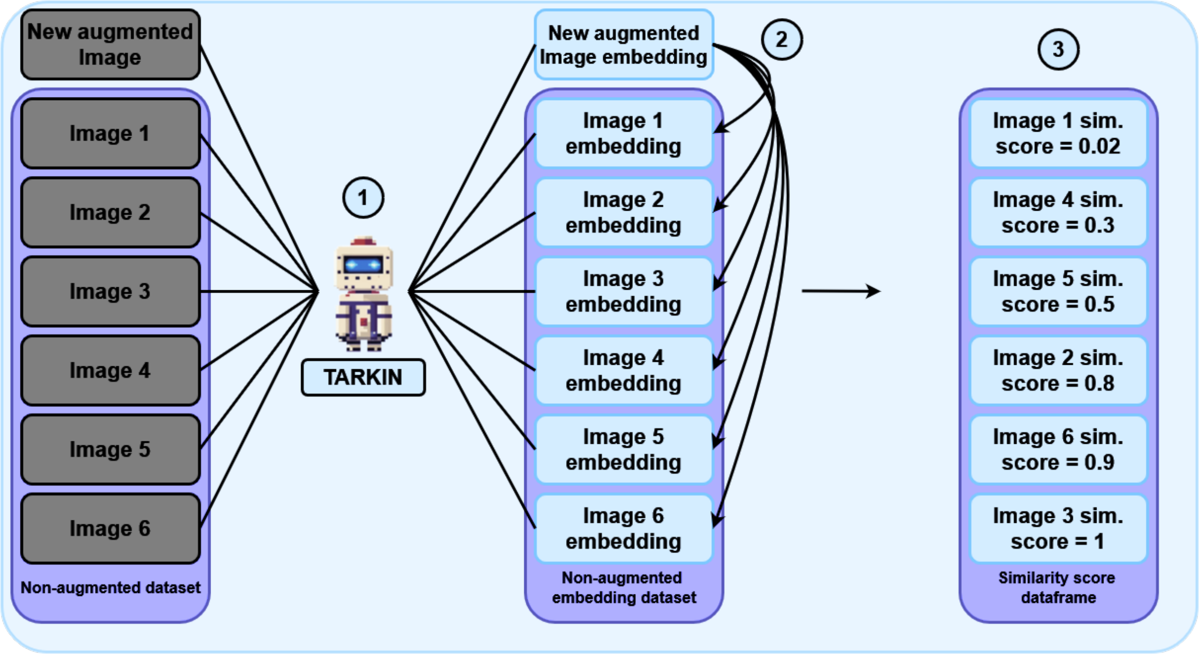


Figure 21: (1) The model (endearingly named TARKIN - Tensor-based Analysis for Redundant Knowledge in Neuroimaging) encodes each image in the MRI dataset into its respective embedding. It also encodes the new image that one wants to find out if it is presently in the dataset. For the accuracy measurement, this image is augmented. (2) The new image's embedding is compared to each image embedding in the dataset by calculating a cosine similarity score for each comparison. (3) The subsequent similarity score dataframe is sorted by the similarity score. If the image name of the image with the highest similarity score is equal to the name of the new image, this counts as a correct match.

In addition, the model was evaluated on if it was capable of retrieving all 365 images from the BraTS-2021 dataset when these were non-augmented and when they were augmented.

2.3.7 Losses

Three types of losses were used for the duplicate recognition task. Firstly, embedding losses that worked on the embedding level. Secondly, reconstruction losses that worked on the input images, and output image reconstructions. Finally combinations of both previous types of losses were also used, named henceforth combination losses.

Only one embedding loss was used during this task. The **triplet margin loss** described in section 4.3.1 by *Equation 8*. Two reconstruction losses were used during this task. The **photometric loss** [53], and the **Structural Similarity Index Measure (SSIM)** loss [54].

The photometric loss is composed of two separate loss functions, the $L1$ loss and the SSIM loss.

The $L1$ loss measures the pixel-wise similarity between the input and the output tensor. It is defined as the absolute error between each pixel of each tensor, and is given by the equation:

$$L1(x,y) = |T_{x,y} - \hat{T}_{x,y}|, \tag{10}$$

where

- $L1(x,y)$ is the L1 loss between the target and predicted values,
- $T_{x,y}$ represents the target values,
- $\hat{T}_{x,y}$ represents the predicted values,
- x, y defines the pixel indices

The SSIM loss measures the similarity between two images, focusing on changes in structural information, luminance, and contrast. It is based on human visual perception, and is given by the equation:

$$\text{SSIM}(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma}\,, \tag{11}$$

where

- x, y are two images of the same size,
- $l(x,y)$ is the luminance comparison function (*Equation 13*),
- $c(x,y)$ is the contrast comparison function (*Equation 14*),
- $s(x,y)$ is the structure comparison function (*Equation 15*),
- α, β, γ are parameters to adjust the relative importance of the three components.

These two losses, the $L1$ loss and the SSIM loss, add together to give the photometric loss used in this experiment. The addition of these loses is subsequently back propagated through the entire network after each iteration of the network.

Finally, two combination losses were also used. The **photometric triplet loss**, a combination of the photometric and the triplet margin loss. Secondly, the **combined loss**, a combination of the SSIM loss and the triplet margin loss. The combinations were done through simple weighted addition, with the photometric loss and SSIM loss having a weight of 0.2, and the triplet margin loss having a weight of 0.8 for each of the combinations. This yielded a total of six losses used for this experiment.

2.4 Results - Duplicate Recognition

During early experimentation with different losses and methods for the MU-NET, an artifact was observed. As seen in *Figure 22 A, B, D*, the reconstruction of the input image generated by the MU-NET contains a checkerboard artifact. Being very noticeable in the SSIM loss and the combined loss, but not very noticeable in the photometric loss. It was first thought that the artifact might be related to certain hyperparameters in the transposed convolutions of the network, especially the stride [55], as is exemplified in *Figure A.25.* Since the MU-NET did not allow for the changing of the transposed convolution parameters, it was decided to use a different implementation of the U-NET from MONAI: BU-NET. As mentioned previously, the BU-NET allowed for the switching of the transposed convolution for simple bi-linear interpolation layers. After switching the transposed convolutional layers for bi-linear interpolation layers, the reconstructions did no longer exhibit the checkerboard artifact, *Figure 22 C.*

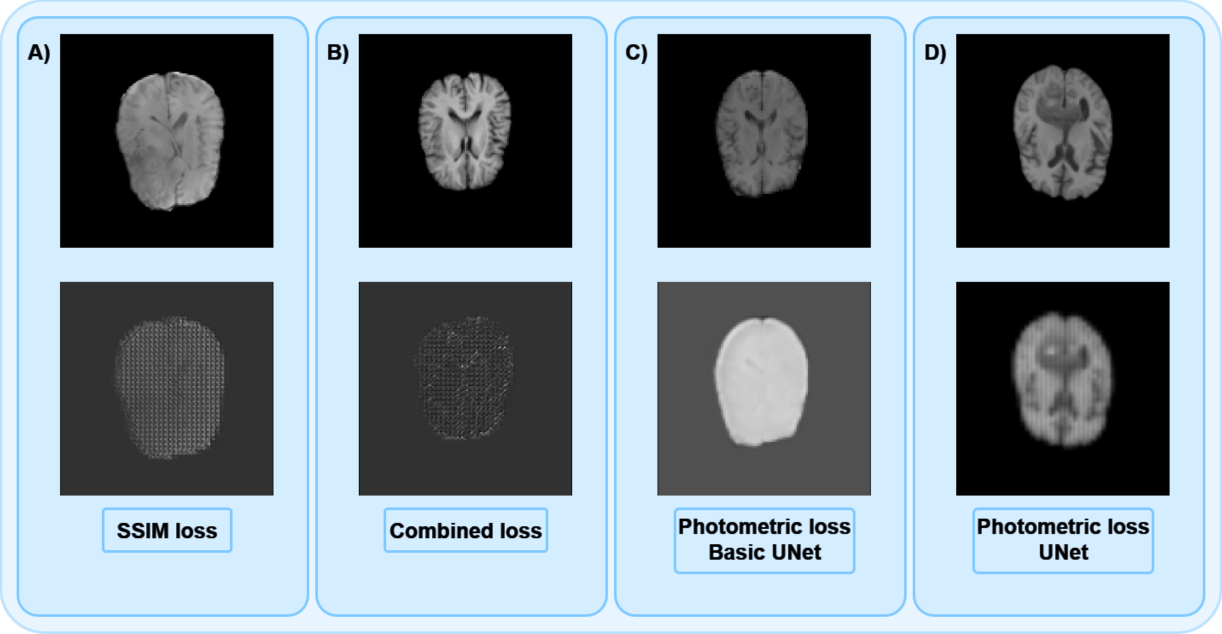


Figure 22: (Top) The original images given to the U-NET architectures. (Bottom) The reconstructions obtained from the networks after 40 epochs of training. Only column (C) corresponds to the BU-NET. Every other column used the MU-NET.

On the MU-NET, the checkerboard artifact also faded when setting the epochs to higher values and training the model for longer using the photometric loss, *Figure 23.*

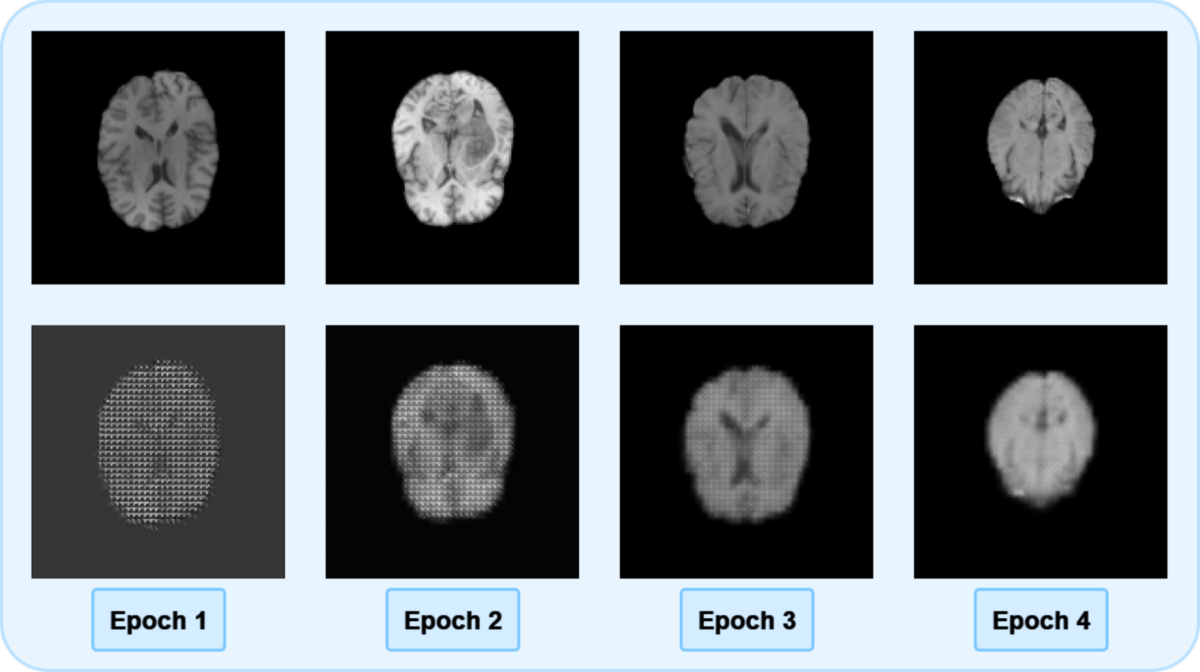


Figure 23: Progressive loss of the checkerboard artifact using the MU-NET architecture and a photometric loss. (Top row) Original images. (Bottom row) Reconstructed images. (Columns) Training epochs, increasing. The checkerboard artifact fades with each training iteration, being very prominent in the first epoch and becoming hardly noticeable towards the later epochs.

Having resolved the checkerboard artifact problem, and other issues that arose early on in this section, the models with different architectures and trained on different losses were compared to each other to determine which one had the highest retrieval accuracy, *Table 5*, *Table 6*, *Table 7*, *Table 8*.

Firstly, the models were trained to identify T1 images from augmented T1 images, as described in Experiment 1. *Table 5* shows how well the models performed for the different augmentation sets.

The accuracy of the models trained on the compact + canonical augmentation set was the lowest. Furthermore, the accuracy improves when the canonical augmentations are removed, *Table 5 column 2*, as it is unable to accurately identify T1 images from T1 images with canonical augmentations. When many augmentations are applied, *Table 5 column 3*, the model performs worse.

Both network architectures performed similarly when using the triplet margin loss, except for when the compact augmentation set was used. Here the SNN model performed better with the triplet margin loss. However, using reconstruction losses proved to be better than simply using embedding losses. Furthermore, there is a notable improvement of the model when over the simple triplet loss when the triplet loss is combined with another loss (combined loss and photometric triplet loss). However, combination losses still perform worse than simple reconstruction losses.

Loss Function	$Accuracy_{T1a}$ Compact + Canonical		$Accuracy_{T1a}$ Compact		$Accuracy_{T1a}$ Extensive	
	UNET	SNN	UNET	SNN	UNET	SNN
Combined Loss	58 %	/	90 %	/	**70 %**	/
Photometric Triplet Loss	59 %	/	89 %	/	**70 %**	/
Triplet Loss	57 %	56 %	89 %	**93 %**	67 %	65 %
SSIM Loss	58 %	/	**95 %**	/	56 %	/
Photometric Loss	**60 %**	/	**95 %**	/	63 %	/

Table 5: Results corresponding to *Experiment 1*. The accuracy of the models in identifying T1 images from augmented T1 images.

An effort was also made to evaluate the models on identifying T1 images from augmented T2 images, as described in Experiment 2. Interestingly, for the reconstruction losses the accuracy of the model was worse when identifying T1 images from augmented T2 images than when identifying T1 images from augmented T1 images. The embedding losses kept the accuracy of the model similar to the previous experiment, which is also true for the photometric triplet loss. However, this time the combined loss performed the best.

Loss Function	$Accuracy_{T2a}$ Compact	
	UNET	**SNN**
Combined Loss	**95 %**	/
Photometric Triplet Loss	89 %	/
Triplet Loss	86 %	**86 %**
SSIM Loss	79 %	/
Photometric Loss	85 %	/

Table 6: Results corresponding to *Experiment 2*. The accuracy of the models in identifying T1 images from augmented T2 images.

Finally, the training done on the dataset containing T1 and T2 augmented images, as described in Experiment 3, yielded the performance results seen in *Table 7*. Here the best performing models were those trained on combination losses.

Loss Function	$Accuracy_{T1/T2a}$ Compact	
	UNET	**SNN**
Combined Loss	**94 %**	/
Photometric Triplet Loss	**94 %**	/
Triplet Loss	92 %	88 %
SSIM Loss	93 %	/
Photometric Loss	92 %	/

Table 7: Results corresponding to *Experiment 2*. The accuracy of the models in identifying T1 images from augmented T1 or T2 images.

Given that the MU-NET SSIM model from the compact augmentations training in the first experiment was the best performing one, it was selected to continue the evaluation of the model. In *Table 8*, the model was evaluated on it's ability to correctly identify all images of the BraTS-2020 dataset (365) that were included in the BraTS-2021 dataset, after the images had been augmented with flips, anisotropies, noise, and a combination of all three. The model performs well when only anisotropic or noise augmentations are applied. It performs worse when flip augmentations are applied.

ID	**Augments**			
	Flip	**Anisotropy**	**Noise**	**Combined**
Correct ID	285	356	356	330
Wrong ID	43	9	9	24
Total	328	365	365	354
Score	0.78	**0.98**	**0.98**	0.90

Table 8: Identification of duplicates in a simulated environment. The score was calculated by dividing the correct identifications by the number of images that the model should have identified (365).

To summarize, during initial experiments with MUNet, a checkerboard artifact appeared in image reconstructions, particularly with SSIM and combined losses. The artifact was traced to the network's transposed convolutions, leading to a switch to BUNet, which uses bilinear interpolation and resolved the issue. Extended training with photometric loss also reduced the artifact. After resolving this, models were compared for accuracy in identifying T1 images from augmented sets. Reconstruction losses outperformed embedding losses with combination losses offering some improvement. In subsequent tests, the MUNet SSIM model performed best overall, particularly in identifying images from the BraTS-2020 dataset in the BraTS-2021 dataset, although it struggled with flip augmentations.

3 Discussion

This section discusses the results obtained for both objectives of this thesis. It goes over the results obtained for the registration, spatial distribution, and duplicate recognition.

3.1 Glioma distribution

3.1.1 Registration

This thesis utilized rigid and deformable registration to register patient's MRIs to several brain atlases with the aim of determining the registration type that would yield the best results given the context of the data at hand. To this end, a wrapper of CaPTk was made to adapt this toolbox to an HPC environment and facilitate its use. This wrapper allowed for the registration of multiple datasets in parallel, in a structured manner, and accounting for different folder structures. In effect, this tool provided a solid base, proved to be invaluable during this thesis, and is a useful output of this thesis that can be used in the future in other projects that require registrations on HPC environments.

For each dataset, the patient images after rigid registration had several problems. The affine registration does not respect the structure of the different brain regions. It tries to simply align the brain with the brain atlas. This has the disadvantage that several structures of the patient MRI will overlap with structures that it shouldn't overlap with in the brain atlas. As can be seen in *Figure 11*, the affine registration of the dataset's tumor distribution overlaps several structures that it shouldn't, such as the brain ventricles *Figure 11 (A)(Middle)* or several brain sulci *Figure 11 (B)(Middle).* This could have caused problems in subsequent analysis, such as determining the severity of tumor attainment of the connectome.

By doing an affine registration, a considerable amount of voxels from the tumor distribution would be mapped to the ventricles in the brain atlas. This would falsify the analysis of the connectome, as there are no white matter tracts in the ventricles that these tumor voxels could attain. The same reasoning would apply to the analysis of the subcortical regions, and cortical regions. The deformable registration does not have these issues. As can be seen in *Figure 11 (B) and (D)*, the spatial distribution does not overlap with the ventricles of the brain atlas nor with the sulci. However, deformable registration does have some disadvantages. For instance, there is no distinction between the brain's regions in MRIs, meaning that the voxels in a specific brain structure of a patient's MRI might be mapped to the wrong structure in the brain atlas.

More so when the brain structures are deformed by a tumor. Still, it is reasonable to assume that more broad, and easily distinguishable features of the brain, such as big sulci, and the ventricles will be correctly mapped, and the tumors mapped to these structures, minimized. Additionally, the distinction between the results after affine, and deformable registration can be quite stark, as can be seen in *Figure 11 (C) and (F)*, with the tumor distribution completely switching location and even the slices demonstrating the highest tumor incidence probability shifting. Still, for the reasons stated beforehand, the analysis for this part of this thesis was done after deformable registration.

An issue that occurred during the registration step was that it became difficult to find a metric that could quantify the quality of the registration. Usually, registration quality relies on determining the overlap between two regions. However, since there were no segmentations of structures in the datasets used for this thesis, this was not feasible [56]. It is also possible to evaluate the registrations by hand and see if the correct brain structures map to the correct place. However, this would have been impractical given the large number of registrations done. Hence, the quality of the registrations was determined based on if the spatial distribution would overlap the ventricles. Similar methods have been applied to gauge registration quality [57]. By this metric, the deformable registration is better than the affine registration.

3.1.2 Spatial distribution

Concerning the spatial distributions of tumors in the different datasets, this thesis found that each dataset's hemispheres are differently attained by tumors, *Figure 16.* BraTS-SSA showed a difference between hemispheres of 1,21%, BraTS-2021 of 5,41%, Burdenko-GBM of 10,99%, LGG of 6,03%, QIN of 32,01%, RHUH-GBM of 17,84%, and UCSF-PDGM of 6,28%. This presents a clear discrepancy with current research, which indicates that there should be no difference between the hemispheres in glioma incidence [58]. One reason for this might be due to laterality errors [59]. Laterality errors, where the MRI image is flipped, are known to happen when composing MRI datasets of extensive sizes. For this thesis, it was difficult to check for the truth of laterality as most available tools, such as AFNI that can detect laterality [60], require modalities that were not available in the datasets used. Hence, the truth of laterality was taken at face value. However, this result could be in part explained by such errors. Other explanations could be that the sample size does not reflect the true underlying population. This is likely to be true for the smaller datasets that have smaller sample sizes, such as QIN. However, BraTS-SSA is also a small dataset and has the smallest difference between hemisphere attainment. This could indicate a higher level curation , and awareness of glioma distribution in the brain, by the team that created this dataset. For larger datasets such as BraTS-2021, it might be a combination of the two aforementioned factors, and others that have not been considered, that lead to this discrepancy.

This thesis also sheds light on a perceived gap in the literature, specifically the lack of data on glioma incidence by anatomical location. One study found that 14% of gliomas occur in deep subcortical regions, while 86% are located in cortical areas, and subcortical white matter [61]. However, further sources to corroborate this are lacking. The results obtained, in *Figure 13* and *Table 2*, give some credence to this claim. The dataset with the lowest incidence of deep subcortical gliomas is LGG, showing only 2.9% of gliomas in deep subcortical regions. In contrast, the BraTS-SSA dataset has the highest incidence, with 17.5% of gliomas occurring in deep subcortical areas. Hence, the proportion of gliomas in the deep subcortical structures seem to somewhat corroborate what has been found in literature. In addition, literature that goes more into depth about the specific deep subcortical structures attained by gliomas was not found. This thesis tries to clarify this point too. *Figure 14* shows to what extent the deep subcortical regions were affected by tumors for each dataset. It is important to clarify that it is not possible to determine the origin of the tumor through the methods employed, only that a tumor is currently in that structure. The identification of glioma origin was outside the goals of this thesis and requires access to specialized machinery, which were not accessible [62]. Still, it was possible to determine that the deep subcortical structures in each dataset were differentially affected by glioma. The structures that showed the most variability between datasets were: the brain stem, the right hippocampus, and the right caudate.

Regarding the cortical regions, by the accounts in current literature, most tumors attain the frontal region, followed by the temporal region, and then the occipital region [61]. As can be seen in *Table 3, Figure 15*, most datasets follow the same order. However some datasets, such as UCSF-PDGM, QIN, BraTS-SSA, have very small differences between all regions. In cases where the dataset is small, this can probably be explained by the small size of the dataset.

Recently the impact of gliomas on the connectome, and the interplay between these two actors has gained increased importance [2]. The impact of a tumor on the survivability, and on the quality of life of a patient has also been noted. A higher attainment of the connectome i.e. white matter tracts, correlates negatively with both aspects. *Figure 17* shows to what extent the connectome is attained in the LGG dataset. In this dataset, 12% of the patients suffer from high connectome attainment, 66% suffer from medium connectome attainment, and 21% suffer from low connectome attainment. In comparison to the other datasets, included in the appendix, it can be said that the patients in this dataset are more likely to have a lower survivability and worse recovery than the patients in the other datasets. Several studies have already tried to utilize MRIs to predict the survival of patients with gliomas to determine care and treatment strategies [63][64]. Furthermore, several challenges, such as the BraTS-2020 challenge, posed to the medical imaging/neuroimaging communities revolve around determining the overall patient survival from MRIs [65]. Given this, a future application of the data generated here, or of the method applied, could be to to inform a model that predicts patient survival from MRI. This might not lead to more performative models, as the model might learn the information about the connectome inadvertently when given enough data, but it could lead to smaller, more efficient, models capable of the same task with the same performance.

A limitation for this section of the thesis was, as mentioned previously, is the uncertainty of if there are any left-right flipped images in the datasets that were analyzed. These errors would have impacted several analysis, especially the distribution between hemispheres. Furthermore, the datasets did not include any information about the patients such as gender or age, which might have been an interesting topic to explore further. Finally, for the connectome part, it would have been best to create a connectome map for each individual patient in each dataset. This would provide a more patient accurate white matter tract distribution, and take into account how the tumor displaces these tracts. However, this would require DWI which the datasets didn't have.

The analysis, done in this thesis, shines light on several aspects of these open-source MRI datasets that might not be considered when first downloading them. The goal of this part of this thesis was to clarify several questions related to the spatial distribution of gliomas in open-source MRI datasets, and uncover aspects that might be useful to inform a DL model or to inform a researcher when selecting a dataset for a DL task. The analysis done regarding the distribution between hemispheres and anatomical locations might aid researchers training a model on generating fake glioma MRIs to augment an existing dataset [66]. Additionally, the analysis done at the connectome level could inform researchers trying to perfect a model to predict the survivability of glioma patients based on MRIs [63]. Moreover, the analysis on the distribution of gliomas between cortical and subcortical structures, can be of value to fill the void in the literature concerning this subject. Also, the tools developed for registration of large datasets are also of importance. Finally, it can be said that the goals envisioned for this part of the thesis were attained and the outputs numerous and fruitful.

3.2 Duplicate Recognition

An interesting find, not pertaining to the main objective of this section, was the checkerboard artifacts generated by the MONAI U-NET implementation. Given that the MONAI framework is a popular framework in the community and its implementations are routinely used, it seems peculiar that there was so little documentation available on this particular problem. It was already known that transposed convolutions might create checkerboard artifacts when the stride of the kernel is not selected to account for this [55]. However, in this case these commonly utilized solutions did not solve the problem. This might indicate an issue with the implementation itself, but this would require further investigation to get to the root of the issue.

As mentioned in the introduction, duplicate images can affect the performance of a DL model [67]. This, combined with the small sizes of glioma MRI datasets, can cause problems for research teams endeavouring to train models on MRI data, for tasks such as tumor augmentation [38], as the data-hungry nature of DL models will inevitably push them to merge publicly available MRI datasets to augment their training data and achieve better performances [68]. To tackle this issue this thesis proposed a model that can detect duplicates of one MRI dataset in another MRI dataset.

The model, named TARKIN for Tensor-based Analysis for Redundant Knowledge in Neuroimaging, was trained using several different embedding, reconstruction, and combination losses. It was noticed that the more distorted, and augmented the images became, the more the model struggled with identifying them correctly, especially when multiple different augmentations were applied. Noticeably, when canonical augmentations or simple flips were applied, the model struggled much more than with any other augmentation. This result was unexpected and can not be confidently explained at the present moment. One supposition would be that, the embedding vectors of an image and its flipped counterpart become mirror images of themselves, and this makes it harder for the cosine similarity function to recognize them as being similar. However, if this were the case, then the SNN should have yielded better results since this architecture has a fully connected neural network at the end, but this is not the case. The SNN performs similar to the U-NET.

The model was trained on 3 different tasks with 3 different datasets. The first was to identify T1 images from augmented T1 images. The second was to identify T1 images from augmented T2 images. The third one was to identify T1 images from augmented T1 or T2 images. Considering only the compact augmentation set and the U-NET architecture, the performance of the model was different in each task, *Table 5, Table 6, Table 7.* The first task yielded the best performing models, and reconstruction losses worked the best. The SSIM & photometric losses, both yielded models with 95% accuracy, while the combination losses, and the embedding loss had lower accuracies. On the second task, combination losses performed better than reconstruction or embedding losses, with the combined loss, and the photometric triplet loss yielding the best models. For the last task the overall accuracy was better, with each loss yielding models that had an accuracy over 90%. These results seem to indicate that for different MRI modalities, a different loss is better suited. Additionally, as mentioned previously, when given T1 images and T2 images the model performs better overall. This might be related to the model seeing more data, and being able to generalize better.

Since, previous, similar studies on this subject have used only a triplet margin loss, an embedding loss, to train these kinds of models [50]. A supplementary objective of this thesis was to clarify what losses are best for duplicate identification. Reasoning being that some losses, such as embedding losses, calculate the loss based on higher dimensional representations of the data, while reconstruction losses might imbue some "biological", or structural, meaning into the embedding vector. This question yielded conflicting answers. The triplet margin loss performs worse than the SSIM, and photometric loss when models are trained on a set of practical augmentations (compact), *Table 5.* However, when more augmentations (extensive) or difficult augmentations (canonical + compact) are applied, the model retains a similar performance or improves. This indicates that the triplet margin loss is a better selection when dealing with a wide variety of augmentations that would strongly distort the image. In addition to this, when merged with a reconstruction loss in a combination loss, the triplet margin loss, reduces the accuracy of the models in most cases, except when the augmentations are extensive. Again this might indicate that, when the images are strongly distorted, that the triplet margin loss has an easier time re-identifying the correct image. In other words, when dealing with a highly abstracted image, a loss that is not dependent on image structure, luminance, similarities, or pixel errors, is better than one that is. In conclusion, reconstruction losses are more useful when the augmentations are lighter and practical, but they perform worse when the augmentations become more and abstract. In that case, a triplet margin loss, or another loss that works on embeddings, should be selected.

When confronted with a test that simulates a real world scenario, *Table 8*, the model proves its value. The model showed to be capable to usefully, detect duplicates in a large sample size. Furthermore, it proved to still be useful when the images were augmented. However, as was already stated before, the model struggles when confronted with flips.

The model seems promising, and could prove to be a useful tool, if some of the issues stated here are ironed out and improved. The cosine similarity measure might not indicated to measure the similarity between embeddings in this case, and it might be beneficial to use another measure for this, such as an euclidean distance or a dot product. However, the best similarity measure to use in this case, might be a similarity measure that is aware of flipped images.

The limitations of this section of the thesis were mostly time based, as this section of the thesis would have benefited from more experimentation with different augmentations and different similarity measures.

4 Outlook: Future Research Directions

Building on the research done in this thesis, several promising avenues for future projects can be envisioned. One significant direction involves leveraging the insights gained from studying the connectome to enhance tumor growth prediction models. As previously discussed, one application of this data could be to predict patient survival rates for individuals diagnosed with gliomas [64]. However, a more nuanced application could involve predicting the direction of tumor growth, an area of considerable interest given the critical role of the tumor microenvironment in determining growth trajectories [69].

The relationship between tumor dynamics and the brain's connectome is well-documented, with numerous studies highlighting the intricate interactions between tumors and neural networks. This includes preferred migratory pathways that tumors tend to follow, which are often guided by the brain's structural connectivity [2][3]. Understanding and predicting these migratory patterns could significantly improve treatment planning and patient outcomes.

A potential approach to address the challenge of predicting tumor growth direction could involve utilizing the connectome atlas developed in this thesis. Specifically, the images from the LUMIERE dataset, a recently established longitudinal glioblastoma MRI dataset [70], could be registered to this connectome atlas. This approach would mirror the methodology employed in this thesis for the BRaTS2020 dataset, where voxel-based analysis was used to integrate white matter tract data.

By weighting each voxel according to the density of white matter tracts passing through it, a predictive model could be trained to incorporate connectomic information. This model would be capable of making more accurate predictions regarding tumor growth direction by taking into account the structural connectivity of the brain. Notably, recent studies have explored similar concepts, demonstrating the potential and pertinence of the suggested approach in the current scientific environment [71].

Pertaining to the model for duplicate image recognition, TARKIN could be improved to become more resistant to canonical images augmentation which would make it even more pertinent and useful. An effort could also be made to incorporate handling of images from different fields. For instance, a logical expansion of the model's capabilities would be to include X-Ray image duplicate to it. X-ray images are substantially more numerous than MRIs and are routinely used in medical assessments and diagnostics, making possible that a patient has multiple X-ray images [72]. Perhaps even some images in different repositories.

Another promising avenue of research for the TARKIN model would be to understand why the model is significantly less accurate when faced with orientation augmented images than other kind of noise. In other words, making the model more interpretable. However this path, has less biological significance and utility, and would be more interesting for someone in the field of interpretable machine learning/artificial intelligence.

In conclusion, this thesis provides a good foundation for several future projects.

References

[1] Krzysztof J. Gorgolewski et al. “The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments”. eng. In: *Scientific Data* 3 (June 2016), p. 160044. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.44.

[2] Alessandro Salvalaggio et al. “Glioblastoma and brain connectivity: the need for a paradigm shift”. eng. In: *The Lancet. Neurology* 23.7 (July 2024), pp. 740–748. ISSN: 1474-4465. DOI: 10.1016/S1474-4422(24)00160-1.

[3] Vishnu Anand Cuddapah et al. “A neurocentric perspective on glioma invasion”. en. In: *Nature Reviews Neuroscience* 15.7 (July 2014). Publisher: Nature Publishing Group, pp. 455–465. ISSN: 1471-0048. DOI: 10.1038/nrn3765. URL: https://www.nature.com/articles/nrn3765 (visited on 08/09/2024).

[4] *What is Neuroimaging? — Psychiatry — U of U School of Medicine.* en. Nov. 2021. URL: https://medicine.utah.edu/psychiatry/research/labs/diagnostic-neuroimaging/neuroimaging (visited on 01/13/2024).

[5] *Scanning the brain.* en. URL: https://www.apa.org/topics/neuropsychology/brain-form-function (visited on 01/10/2024).

[6] admin. *Difference Between an MRI and an X-Ray — SI Ortho.* en-US. Dec. 2022. URL: https://siortho.com/blog/x-rays/difference-between-mri-xray/ (visited on 01/13/2024).

[7] *Magnetic Resonance in Medicine, e-Textbook by Peter A. Rinck.* URL: http://magnetic-resonance.org/ (visited on 01/13/2024).

[8] Katarzyna Krupa and Monika Bekiesińska-Figatowska. “Artifacts in Magnetic Resonance Imaging”. In: *Polish Journal of Radiology* 80 (Feb. 2015), pp. 93–106. ISSN: 1733-134X. DOI: 10.12659/PJR.892628. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4340093/ (visited on 08/06/2024).

[9] Martin J. Graves and Donald G. Mitchell. “Body MRI artifacts in clinical practice: a physicist’s and radiologist’s perspective”. eng. In: *Journal of magnetic resonance imaging: JMRI* 38.2 (Aug. 2013), pp. 269–287. ISSN: 1522-2586. DOI: 10.1002/jmri.24288.

[10] D. Bell, Y. Glick, and M. Thurston. “Reference article, Radiopedia.org”. In: (Apr. 2022). DOI: https://doi.org/10.53347/rID-61013. URL: https://radiopaedia.org/articles/61013.

[11] Vladimir Fonov et al. "Unbiased average age-appropriate atlases for pediatric studies". In: *NeuroImage* 54.1 (Jan. 2011), pp. 313–327. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2010.07.033. URL: https://www.sciencedirect.com/science/article/pii/S1053811910010062 (visited on 08/15/2024).

[12] Torsten Rohlfing et al. "The SRI24 multichannel atlas of normal adult human brain structure". eng. In: *Human Brain Mapping* 31.5 (May 2010), pp. 798–819. ISSN: 1097-0193. DOI: 10.1002/hbm.20906.

[13] *Gliomas*. en. Feb. 2022. URL: https://www.hopkinsmedicine.org/health/conditions-and-diseases/gliomas (visited on 01/13/2024).

[14] McKinsey L. Goodenberger and Robert B. Jenkins. "Genetics of adult glioma". English. In: *Cancer Genetics* 205.12 (Dec. 2012). Publisher: Elsevier, pp. 613–621. ISSN: 2210-7762. DOI: 10.1016/j.cancergen.2012.10.009. URL: https://www.cancergeneticsjournal.org/article/S2210-7762(12)00260-8/fulltext (visited on 01/13/2024).

[15] Elizabeth A. Maher et al. "Malignant glioma: genetics and biology of a grave matter". en. In: *Genes & Development* 15.11 (June 2001). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1311–1333. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.891601. URL: http://genesdev.cshlp.org/content/15/11/1311 (visited on 01/13/2024).

[16] *Glioma - Symptoms, Causes, Treatment — NORD*. en-US. URL: https://rarediseases.org/rare-diseases/glioma/ (visited on 01/13/2024).

[17] A. Lim et al. "Complex visual hallucinations as a presentation of temporal low-grade glioma". English. In: *Journal of Clinical Neuroscience* 18.1 (Jan. 2011). Publisher: Elsevier, pp. 157–159. ISSN: 0967-5868, 1532-2653. DOI: 10.1016/j.jocn.2010.07.112. URL: https://www.jocn-journal.com/article/S0967-5868(10)00647-8/fulltext (visited on 01/13/2024).

[18] Nicolas R Smoll et al. "Computed tomography scan radiation and brain cancer incidence". In: *Neuro-Oncology* 25.7 (July 2023), pp. 1368–1376. ISSN: 1522-8517. DOI: 10.1093/neuonc/noad012. URL: https://doi.org/10.1093/neuonc/noad012 (visited on 01/13/2024).

[19] Kaveh Barami. "Oncomodulatory mechanisms of human cytomegalovirus in gliomas". English. In: *Journal of Clinical Neuroscience* 17.7 (July 2010). Publisher: Elsevier, pp. 819–823. ISSN: 0967-5868, 1532-2653. DOI: 10.1016/j.jocn.2009.10.040. URL: https:

//www.jocn-journal.com/article/S0967-5868(10)00074-3/fulltext (visited on 01/13/2024).

[20] Maral Adel Fahmideh et al. “Association between DNA repair gene polymorphisms and risk of glioma: A systematic review and meta-analysis”. In: *Neuro-Oncology* 16.6 (June 2014), pp. 807–814. ISSN: 1522-8517. DOI: 10.1093/neuonc/nou003. URL: https://doi.org/10.1093/neuonc/nou003 (visited on 01/13/2024).

[21] Remco J. Molenaar et al. “The driver and passenger effects of isocitrate dehydrogenase 1 and 2 mutations in oncogenesis and survival prolongation”. In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1846.2 (Dec. 2014), pp. 326–341. ISSN: 0304-419X. DOI: 10.1016/j.bbcan.2014.05.004. URL: https://www.sciencedirect.com/science/article/pii/S0304419X14000493 (visited on 01/13/2024).

[22] David N. Louis et al. “The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary”. en. In: *Acta Neuropathologica* 131.6 (June 2016), pp. 803–820. ISSN: 1432-0533. DOI: 10.1007/s00401-016-1545-1. URL: https://doi.org/10.1007/s00401-016-1545-1 (visited on 01/13/2024).

[23] Olaf Sporns, Giulio Tononi, and Rolf Kötter. “The Human Connectome: A Structural Description of the Human Brain”. In: *PLoS Computational Biology* 1.4 (Sept. 2005), e42. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.0010042. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1239902/ (visited on 08/06/2024).

[24] Valéria Pereira Ferrer, Vivaldo Moura Neto, and Rolf Mentlein. “Glioma infiltration and extracellular matrix: key players and modulators”. eng. In: *Glia* 66.8 (Aug. 2018), pp. 1542–1565. ISSN: 1098-1136. DOI: 10.1002/glia.23309.

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. en. In: *Nature* 521.7553 (May 2015). Publisher: Nature Publishing Group, pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539. URL: https://www.nature.com/articles/nature14539 (visited on 08/06/2024).

[26] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. en. New York, NY: Springer New York, 2013. ISBN: 978-1-4614-6848-6 978-1-4614-6849-3. DOI: 10.1007/978-1-4614-6849-3. URL: http://link.springer.com/10.1007/978-1-4614-6849-3 (visited on 08/07/2024).

[27] Rikiya Yamashita et al. “Convolutional neural networks: an overview and application in radiology”. en. In: *Insights into Imaging* 9.4 (Aug. 2018). Number: 4 Publisher: SpringerOpen, pp. 611–629. ISSN: 1869-4101. DOI: 10.1007/s13244-018-0639-9. URL: https://

insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9 (visited on 08/06/2024).

[28] Shun-ichi Amari. “Backpropagation and stochastic gradient descent method”. In: *Neurocomputing* 5.4 (June 1993), pp. 185–196. ISSN: 0925-2312. DOI: 10.1016/0925-2312(93)90006-O. URL: https://www.sciencedirect.com/science/article/pii/092523129390006O (visited on 08/06/2024).

[29] S. Postalcioglu. “Performance Analysis of Different Optimizers for Deep Learning-Based Image Recognition”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 34 (Apr. 2019). DOI: 10.1142/S0218001420510039.

[30] Brian D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge: Cambridge University Press, 1996. ISBN: 978-0-521-71770-0. DOI: 10.1017/CBO9780511812651. URL: https://www.cambridge.org/core/books/pattern-recognition-and-neural-networks/4E038249C9BAA06C8F4EE6F044D09C5C (visited on 08/06/2024).

[31] Andreea Roxana Luca et al. “Impact of quality, type and volume of data used by deep learning models in the analysis of medical images”. In: *Informatics in Medicine Unlocked* 29 (Jan. 2022), p. 100911. ISSN: 2352-9148. DOI: 10.1016/j.imu.2022.100911. URL: https://www.sciencedirect.com/science/article/pii/S2352914822000612 (visited on 08/06/2024).

[32] Harini Suresh and John V. Guttag. “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle”. In: *Equity and Access in Algorithms, Mechanisms, and Optimization.* arXiv:1901.10002 [cs, stat]. Oct. 2021, pp. 1–9. DOI: 10.1145/3465416.3483305. URL: http://arxiv.org/abs/1901.10002 (visited on 08/06/2024).

[33] Matteo Renzulli et al. “Segmental Distribution of Hepatocellular Carcinoma in Cirrhotic Livers”. In: *Diagnostics* 12.4 (Mar. 2022), p. 834. ISSN: 2075-4418. DOI: 10.3390/diagnostics12040834. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9032124/ (visited on 08/06/2024).

[34] Raphael Maser et al. “Unified Retrieval for Streamlining Biomedical Image Dataset Aggregation and Standardization”. de. In: *Bildverarbeitung für die Medizin 2024.* Ed. by Andreas Maier et al. Wiesbaden: Springer Fachmedien, 2024, pp. 328–333. ISBN: 978-3-658-44037-4. DOI: 10.1007/978-3-658-44037-4_83.

[35] Yangming Ou et al. “Comparative Evaluation of Registration Algorithms in Different Brain Databases With Varying Difficulty: Results and Insights”. en. In: *IEEE Transactions on Medical Imaging* 33.10 (Oct. 2014), pp. 2039–2065. ISSN: 0278-0062, 1558-254X. DOI:

10.1109/TMI.2014.2330355. URL: http://ieeexplore.ieee.org/document/6834815/ (visited on 08/07/2024).

[36] Mahsa Dadar et al. “MNI-FTD templates, unbiased average templates of frontotemporal dementia variants”. en. In: *Scientific Data* 8.1 (Aug. 2021). Publisher: Nature Publishing Group, p. 222. ISSN: 2052-4463. DOI: 10.1038/s41597-021-01007-5. URL: https://www.nature.com/articles/s41597-021-01007-5 (visited on 08/07/2024).

[37] Sarthak Pati et al. “The Cancer Imaging Phenomics Toolkit (CaPTk): Technical Overview”. eng. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes (Workshop)* 11993 (2020), pp. 380–394. DOI: 10.1007/978-3-030-46643-5_38.

[38] Konstantinos Kamnitsas et al. “DeepMedic for Brain Tumor Segmentation”. In: Apr. 2016, pp. 138–149. ISBN: 978-3-319-55523-2. DOI: 10.1007/978-3-319-55524-9_14.

[39] Paul Yushkevich et al. “FAST AUTOMATIC SEGMENTATION OF HIPPOCAMPAL SUBFIELDS AND MEDIAL TEMPORAL LOBE SUBREGIONS IN 3 TESLA AND 7 TESLA T2-WEIGHTED MRI”. In: *Alzheimer’s & Dementia* 12 (July 2016), P126–P127. DOI: 10.1016/j.jalz.2016.06.205.

[40] Brian B Avants, Nick Tustison, and Hans Johnson. “Advanced Normalization Tools (ANTS)”. en. In: ().

[41] Q. R. Razlighi, N. Kehtarnavaz, and S. Yousefi. “Evaluating Similarity Measures for Brain Image Registration”. In: *Journal of visual communication and image representation* 24.7 (Oct. 2013), pp. 977–987. ISSN: 1047-3203. DOI: 10.1016/j.jvcir.2013.06.010. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3771653/ (visited on 08/06/2024).

[42] Yasser Alemán-Gómez et al. “A multi-scale probabilistic atlas of the human connectome”. en. In: *Scientific Data* 9.1 (Aug. 2022). Publisher: Nature Publishing Group, p. 516. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01624-8. URL: https://www.nature.com/articles/s41597-022-01624-8 (visited on 08/06/2024).

[43] Michel Bilello et al. “Population-based MRI atlases of spatial distribution are specific to patient and tumor characteristics in glioblastoma”. eng. In: *NeuroImage. Clinical* 12 (2016), pp. 34–40. ISSN: 2213-1582. DOI: 10.1016/j.nicl.2016.03.007.

[44] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. “Siamese Neural Networks for One-shot Image Recognition”. en. In: ().

[45] Abhilash Nandy et al. “A Survey on Applications of Siamese Neural Networks in Computer Vision”. In: June 2020, pp. 1–5. DOI: 10.1109/INCET49848.2020.9153977.

[46] Vassileios Balntas et al. “Learning local feature descriptors with triplets and shallow convolutional neural networks”. In: Jan. 2016, pp. 119.1–119.11. DOI: 10.5244/C.30.119.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: vol. 9351. Oct. 2015, pp. 234–241. ISBN: 978-3-319-24573-7. DOI: 10.1007/978-3-319-24574-4_28.

[48] Xiaodan Hu et al. “RUNet: A Robust UNet Architecture for Image Super-Resolution”. en. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Long Beach, CA, USA: IEEE, June 2019, pp. 505–507. ISBN: 978-1-72812-506-0. DOI: 10.1109/CVPRW.2019.00073. URL: https://ieeexplore.ieee.org/document/9025499/ (visited on 08/06/2024).

[49] Yanyu Li et al. *SnapFusion: Text-to-Image Diffusion Model on Mobile Devices within Two Seconds.* arXiv:2306.00980 [cs]. Oct. 2023. DOI: 10.48550/arXiv.2306.00980. URL: http://arxiv.org/abs/2306.00980 (visited on 08/06/2024).

[50] Kai Packhäuser et al. “Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data”. en. In: *Scientific Reports* 12.1 (Sept. 2022). Publisher: Nature Publishing Group, p. 14851. ISSN: 2045-2322. DOI: 10.1038/s41598-022-19045-3. URL: https://www.nature.com/articles/s41598-022-19045-3 (visited on 08/06/2024).

[51] *For academic research.* en-US. URL: https://wandb.ai/site/research (visited on 08/15/2024).

[52] Henning Müller et al. “Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals”. In: *Pattern Recognition Letters* (Apr. 2001), pp. 593–601. DOI: 10.1016/S0167-8655(00)00118-5.

[53] Clement Godard et al. “Digging Into Self-Supervised Monocular Depth Estimation”. en. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV).* Seoul, Korea (South): IEEE, Oct. 2019, pp. 3827–3837. ISBN: 978-1-72814-803-8. DOI: 10.1109/ICCV.2019.00393. URL: https://ieeexplore.ieee.org/document/9009796/ (visited on 08/13/2024).

[54] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4 (Apr. 2004). Conference Name: IEEE Transactions on Image Processing, pp. 600–612. ISSN: 1941-0042. DOI: 10.1109/TIP.2003.819861. URL: https://ieeexplore.ieee.org/document/1284395 (visited on 08/13/2024).

[55] Augustus Odena, Vincent Dumoulin, and Chris Olah. "Deconvolution and Checkerboard Artifacts". en. In: *Distill* 1.10 (Oct. 2016), e3. ISSN: 2476-0757. DOI: 10.23915/distill.00003. URL: http://distill.pub/2016/deconv-checkerboard (visited on 08/15/2024).

[56] Yi Rong et al. "Rigid and Deformable Image Registration for Radiation Therapy: A Self-Study Evaluation Guide for NRG Oncology Clinical Trial Participation". In: *Practical radiation oncology* 11.4 (2021), pp. 282–298. ISSN: 1879-8500. DOI: 10.1016/j.prro.2021.02.007. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8406084/ (visited on 08/15/2024).

[57] Florian Dubost et al. "Multi-atlas image registration of clinical data with automated quality assessment using ventricle segmentation". In: *Medical image analysis* 63 (July 2020), p. 101698. ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101698. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7275913/ (visited on 08/06/2024).

[58] Juan Jose Valenzuela-Fuenzalida et al. "Association between the Anatomical Location of Glioblastoma and Its Evaluation with Clinical Considerations: A Systematic Review and Meta-Analysis". en. In: *Journal of Clinical Medicine* 13.12 (Jan. 2024). Number: 12 Publisher: Multidisciplinary Digital Publishing Institute, p. 3460. ISSN: 2077-0383. DOI: 10.3390/jcm13123460. URL: https://www.mdpi.com/2077-0383/13/12/3460 (visited on 08/06/2024).

[59] Subba Rao Digumarthy et al. "Process improvement for reducing side discrepancies in radiology reports". en. In: *Acta Radiologica Open* 7.7-8 (Aug. 2018), p. 205846011879472. ISSN: 2058-4601, 2058-4601. DOI: 10.1177/2058460118794727. URL: http://journals.sagepub.com/doi/10.1177/2058460118794727 (visited on 08/15/2024).

[60] Daniel R. Glen et al. "Beware (Surprisingly Common) Left-Right Flips in Your MRI Data: An Efficient and Robust Method to Check MRI Dataset Consistency Using AFNI". English. In: *Frontiers in Neuroinformatics* 0 (May 2020). Publisher: Frontiers. ISSN: 1662-5196. DOI: 10.3389/fninf.2020.00018. URL: https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2020.00018/full (visited on 08/06/2024).

[61] Suvi Larjavaara et al. "Incidence of gliomas by anatomic location". In: *Neuro-Oncology* 9.3 (July 2007), pp. 319–325. ISSN: 1522-8517. DOI: 10.1215/15228517-2007-016. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1907421/ (visited on 08/06/2024).

[62] Sheila R Alcantara Llaguno and Luis F Parada. "Cell of origin of glioma: biological and clinical implications". In: *British Journal of Cancer* 115.12 (Dec. 2016), pp. 1445–1450.

ISSN: 0007-0920. DOI: 10.1038/bjc.2016.354. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5155355/ (visited on 08/06/2024).

[63] Samin Babaei Rikan et al. "Survival prediction of glioblastoma patients using modern deep learning and machine learning techniques". en. In: *Scientific Reports* 14.1 (Jan. 2024), p. 2371. ISSN: 2045-2322. DOI: 10.1038/s41598-024-53006-2. URL: https://www.nature.com/articles/s41598-024-53006-2 (visited on 08/15/2024).

[64] Sveinn Pálsson et al. "Predicting survival of glioblastoma from automatic whole-brain and tumor segmentation of MR images". en. In: *Scientific Reports* 12.1 (Nov. 2022). Publisher: Nature Publishing Group, p. 19744. ISSN: 2045-2322. DOI: 10.1038/s41598-022-19223-3. URL: https://www.nature.com/articles/s41598-022-19223-3 (visited on 08/06/2024).

[65] Bjoern H. Menze et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)". eng. In: *IEEE transactions on medical imaging* 34.10 (Oct. 2015), pp. 1993–2024. ISSN: 1558-254X. DOI: 10.1109/TMI.2014.2377694.

[66] Debadyuti Mukherkjee et al. "Brain tumor image generation using an aggregation of GAN models with style transfer". en. In: *Scientific Reports* 12.1 (June 2022). Publisher: Nature Publishing Group, p. 9141. ISSN: 2045-2322. DOI: 10.1038/s41598-022-12646-y. URL: https://www.nature.com/articles/s41598-022-12646-y (visited on 08/06/2024).

[67] L. Manjusha and V. Suryanarayana. "Detect /Remove Duplicate Images from a Dataset for Deep Learning". en. In: *Journal of Positive School Psychology* 6.2 (Mar. 2022). Number: 2, pp. 606–609. ISSN: 2717-7564. URL: https://journalppw.com/index.php/jpsp/article/view/1547 (visited on 08/06/2024).

[68] Gary Marcus. *Deep Learning: A Critical Appraisal.* arXiv:1801.00631 [cs, stat]. Jan. 2018. DOI: 10.48550/arXiv.1801.00631. URL: http://arxiv.org/abs/1801.00631 (visited on 08/06/2024).

[69] Yixun Liu et al. "Patient specific tumor growth prediction using multimodal images". In: *Medical Image Analysis* 18.3 (Apr. 2014), pp. 555–566. ISSN: 1361-8415. DOI: 10.1016/j.media.2014.02.005. URL: https://www.sciencedirect.com/science/article/pii/S1361841514000280 (visited on 08/09/2024).

[70] Yannick Suter et al. "The LUMIERE dataset: Longitudinal Glioblastoma MRI with expert RANO evaluation". In: *Scientific Data* 9 (Dec. 2022), p. 768. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01881-7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9755255/ (visited on 08/09/2024).

[71] Michel Friedrich et al. "Structural connectome-based predictive modeling of cognitive deficits in treated glioma patients". In: *Neuro-Oncology Advances* 6.1 (Jan. 2024), vdad151. ISSN: 2632-2498. DOI: `10.1093/noajnl/vdad151`. URL: `https://doi.org/10.1093/noajnl/vdad151` (visited on 08/09/2024).

[72] Martin Spahn. "X-ray detectors in medical imaging". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. PIXEL 2012 731 (Dec. 2013), pp. 57–63. ISSN: 0168-9002. DOI: `10.1016/j.nima.2013.05.174`. URL: `https://www.sciencedirect.com/science/article/pii/S0168900213007961` (visited on 08/09/2024).

A Appendix

All the code used for this thesis is available at: https://github.com/Bananahopper

A.1 Supplementary glioma figures

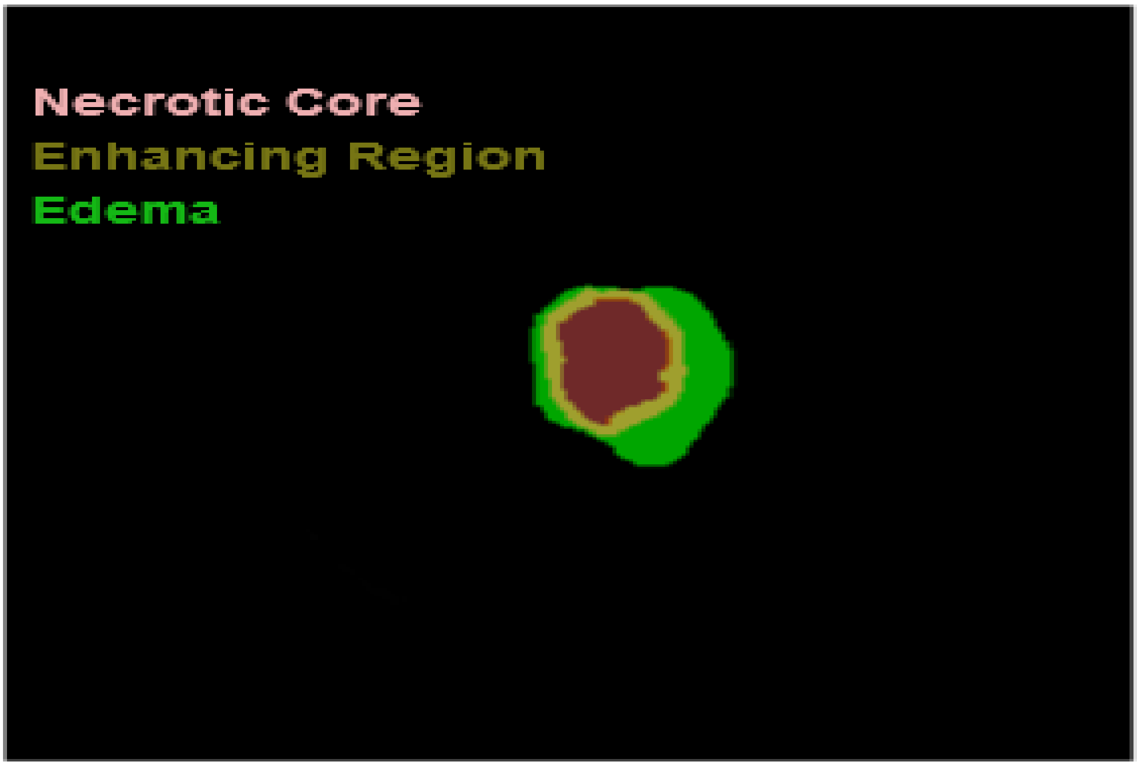


Figure A.1: An example of a tumor segmentation as seen in MRI. Coloring corresponds to a specific label

A.2 Supplementary machine learning figures and concepts

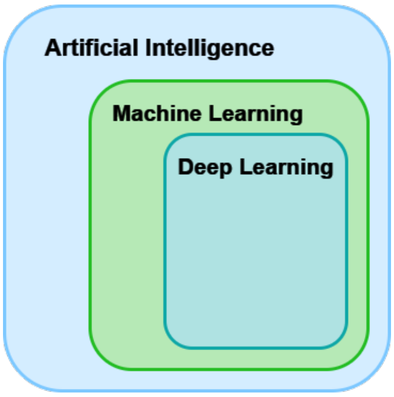


Figure A.2: Artificial Intelligence is a research field that focuses on sensing, reasoning, adapting, and planning algorithms. Machine Learning refers to a research field that investigates algorithms that can find patterns in data, and improve their performance as they are exposed to more data over time. DL refers to a research field that specifically uses multilayered neural networks that learn from vast amounts of data.

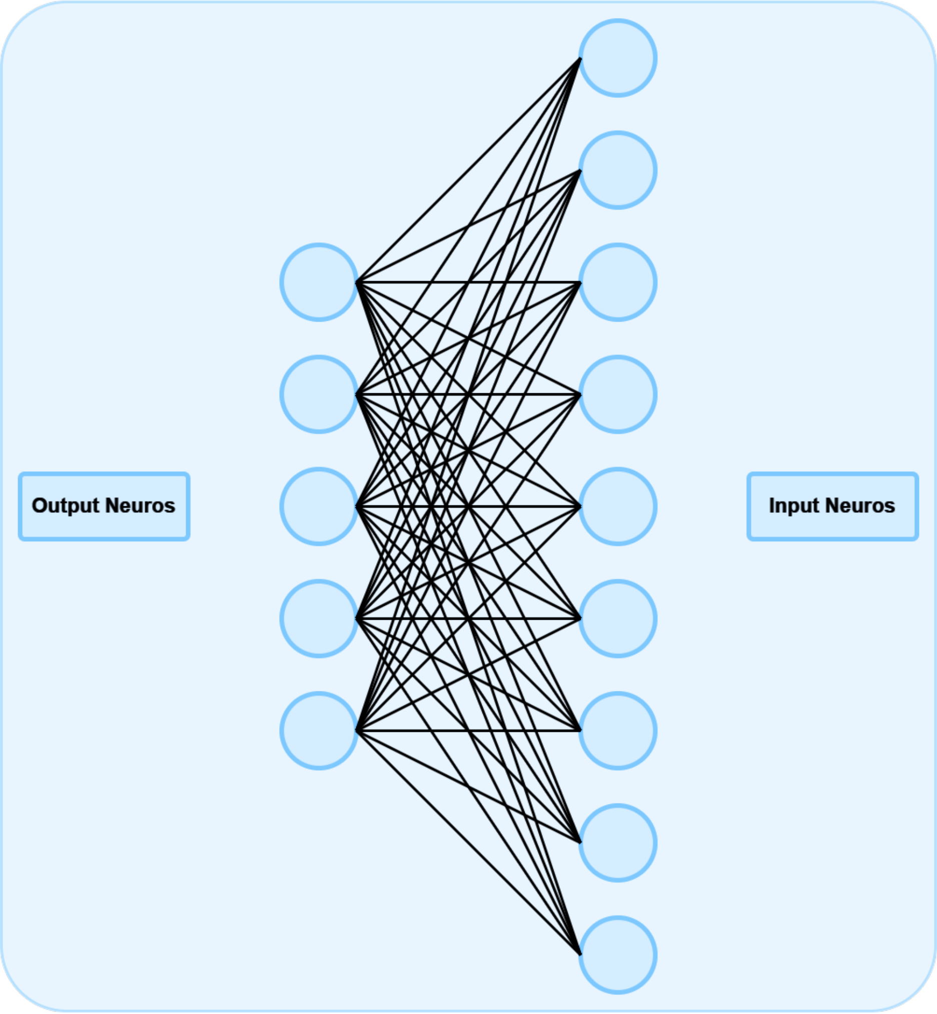


Figure A.3: The fully connected layer. (Left) The input features. (Right) The output features. Each input feature is connected to each output feature, by *Equation 12*.

$$y_{jk}(x) = f\left(\sum_{i=1}^{nH} (w_{jk} \cdot x_i + w_{j0})\right), \tag{12}$$

A.3 Supplementary glioma distribution plots and figures

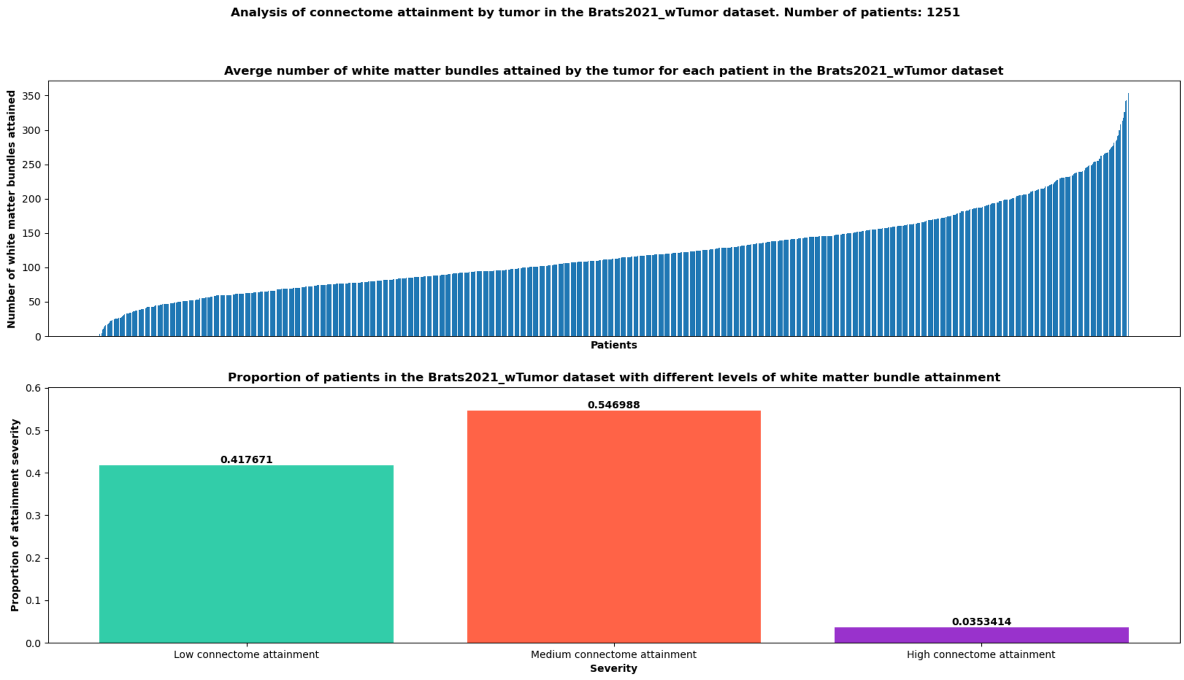


Figure A.4: (Top) The overlap between the tumor segmentations & the connectome white matter tracts. (Bottom) The percentage of low, medium, and high connectome attainment in the dataset. Low attainment = average white matter tracts attained by the tumor < 100. Medium attainment = $100 <$ average white matter tracts attained by the tumor < 250. High attainment = average white matter tracts attained by the tumor > 250.

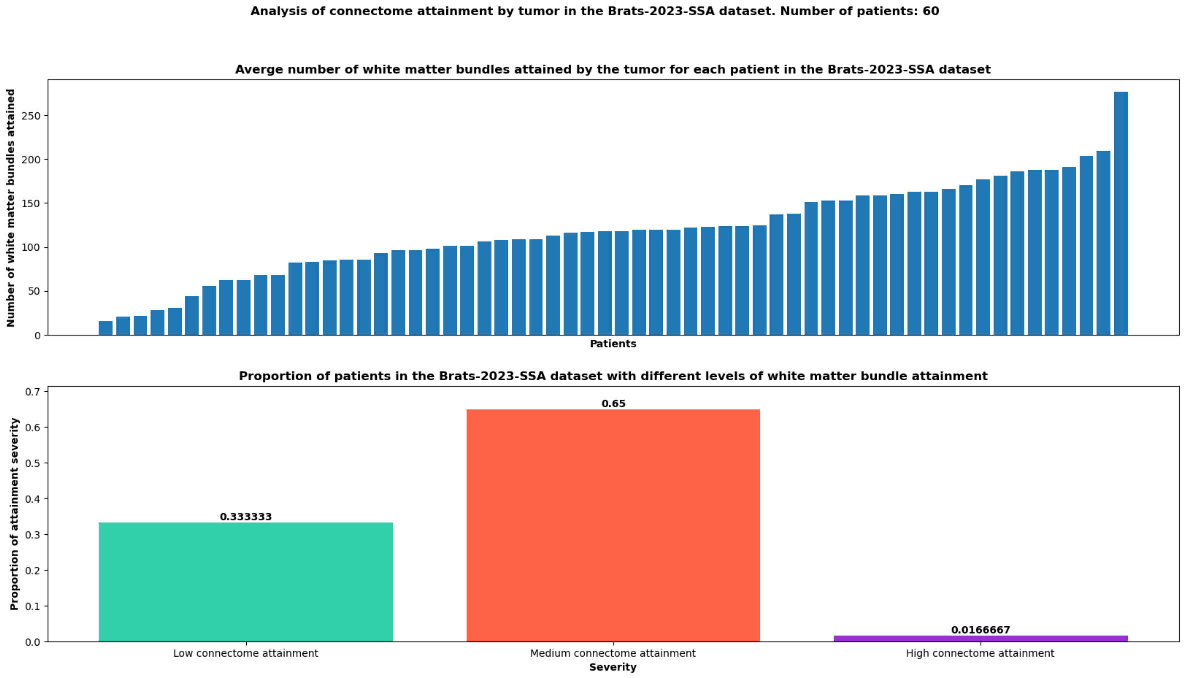


Figure A.5: (Top) The overlap between the tumor segmentations & the connectome white matter tracts. (Bottom) The percentage of low, medium, and high connectome attainment in the dataset. Low attainment = average white matter tracts attained by the tumor < 100. Medium attainment = 100 < average white matter tracts attained by the tumor < 250. High attainment = average white matter tracts attained by the tumor > 250.

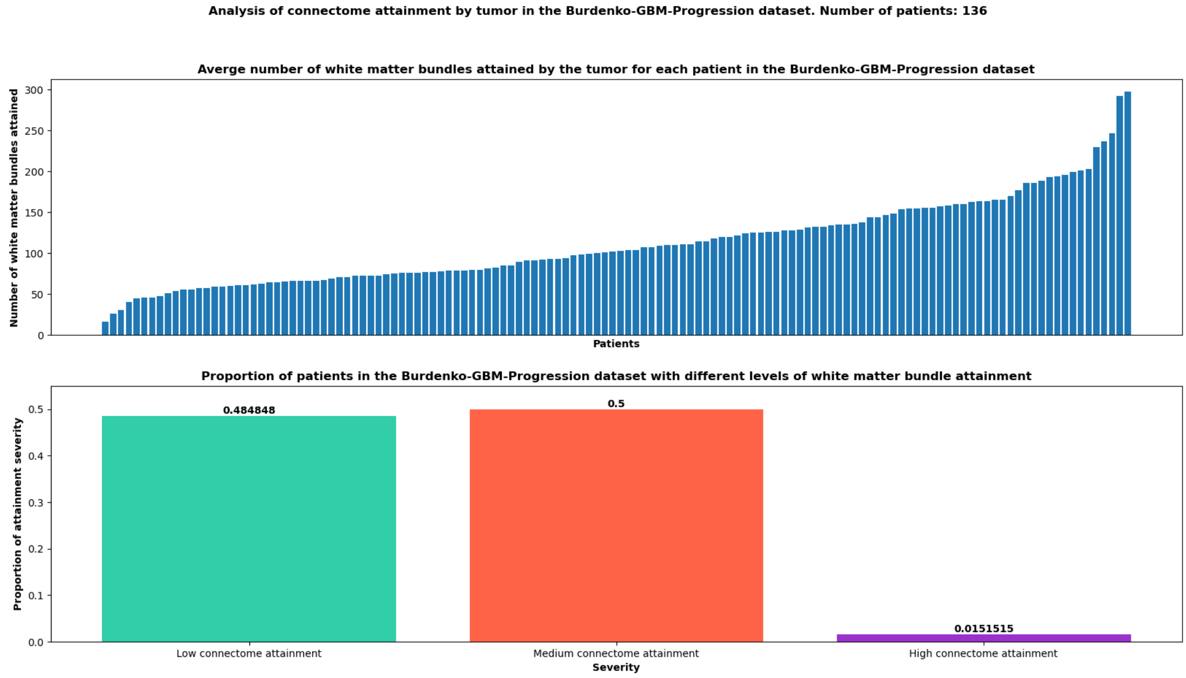


Figure A.6: (Top) The overlap between the tumor segmentations & the connectome white matter tracts. (Bottom) The percentage of low, medium, and high connectome attainment in the dataset. Low attainment = average white matter tracts attained by the tumor < 100. Medium attainment = 100 < average white matter tracts attained by the tumor < 250. High attainment = average white matter tracts attained by the tumor > 250.

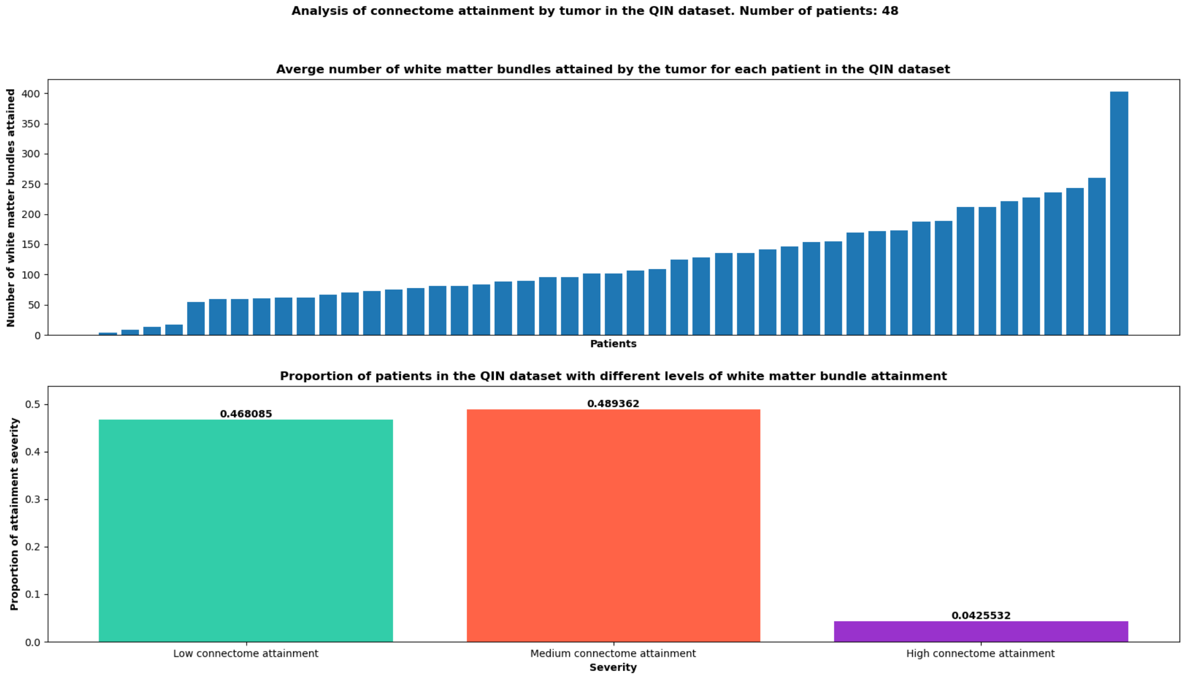


Figure A.7: (Top) The overlap between the tumor segmentations & the connectome white matter tracts. (Bottom) The percentage of low, medium, and high connectome attainment in the dataset. Low attainment = average white matter tracts attained by the tumor < 100. Medium attainment = 100 < average white matter tracts attained by the tumor < 250. High attainment = average white matter tracts attained by the tumor > 250.

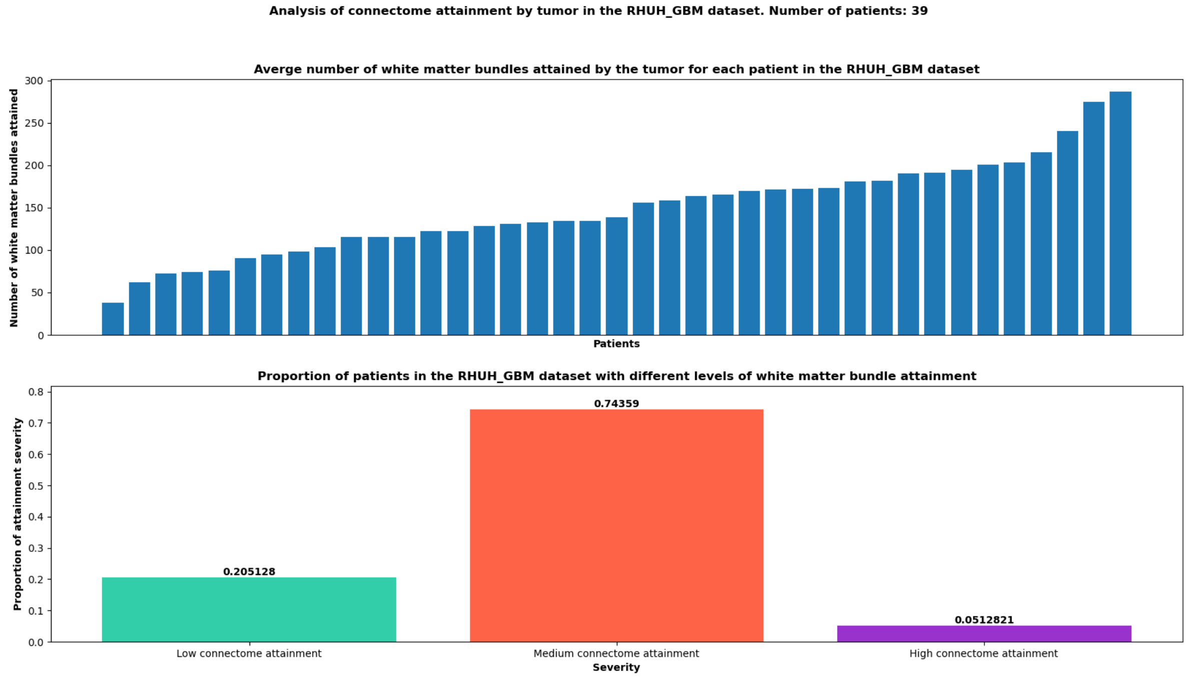


Figure A.8: (Top) The overlap between the tumor segmentations & the connectome white matter tracts. (Bottom) The percentage of low, medium, and high connectome attainment in the dataset. Low attainment = average white matter tracts attained by the tumor < 100. Medium attainment = $100 <$ average white matter tracts attained by the tumor < 250. High attainment = average white matter tracts attained by the tumor > 250.

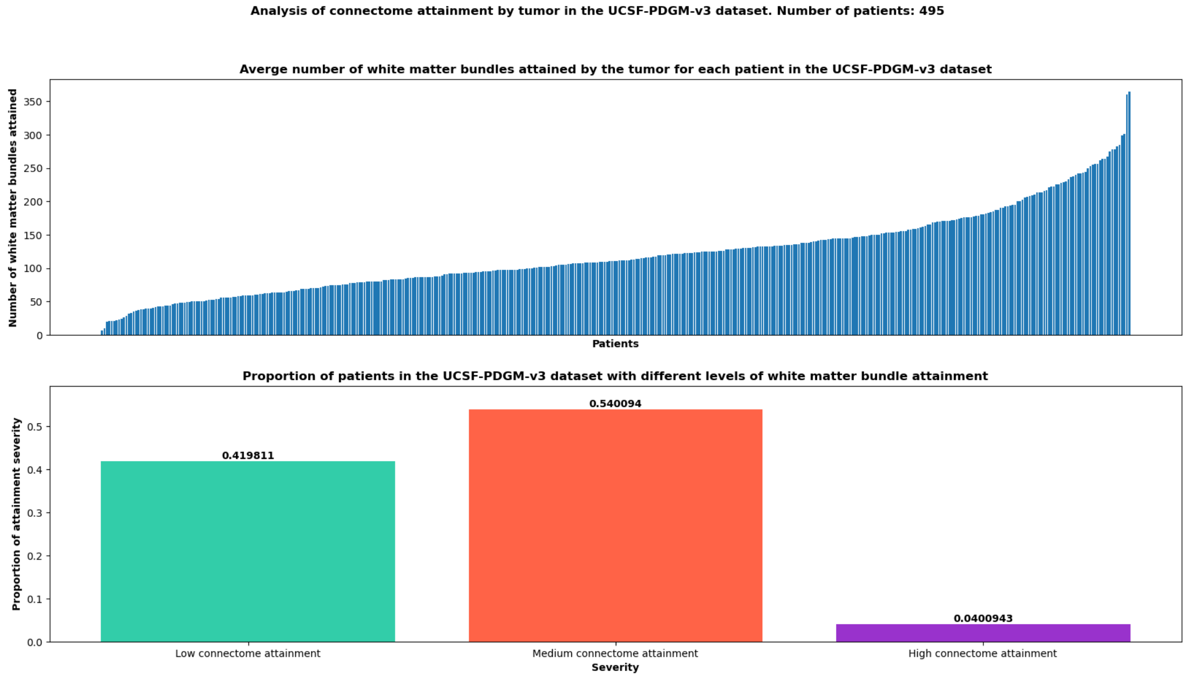


Figure A.9: (Top) The overlap between the tumor segmentations & the connectome white matter tracts. (Bottom) The percentage of low, medium, and high connectome attainment in the dataset. Low attainment = average white matter tracts attained by the tumor < 100. Medium attainment = 100 < average white matter tracts attained by the tumor < 250. High attainment = average white matter tracts attained by the tumor > 250.

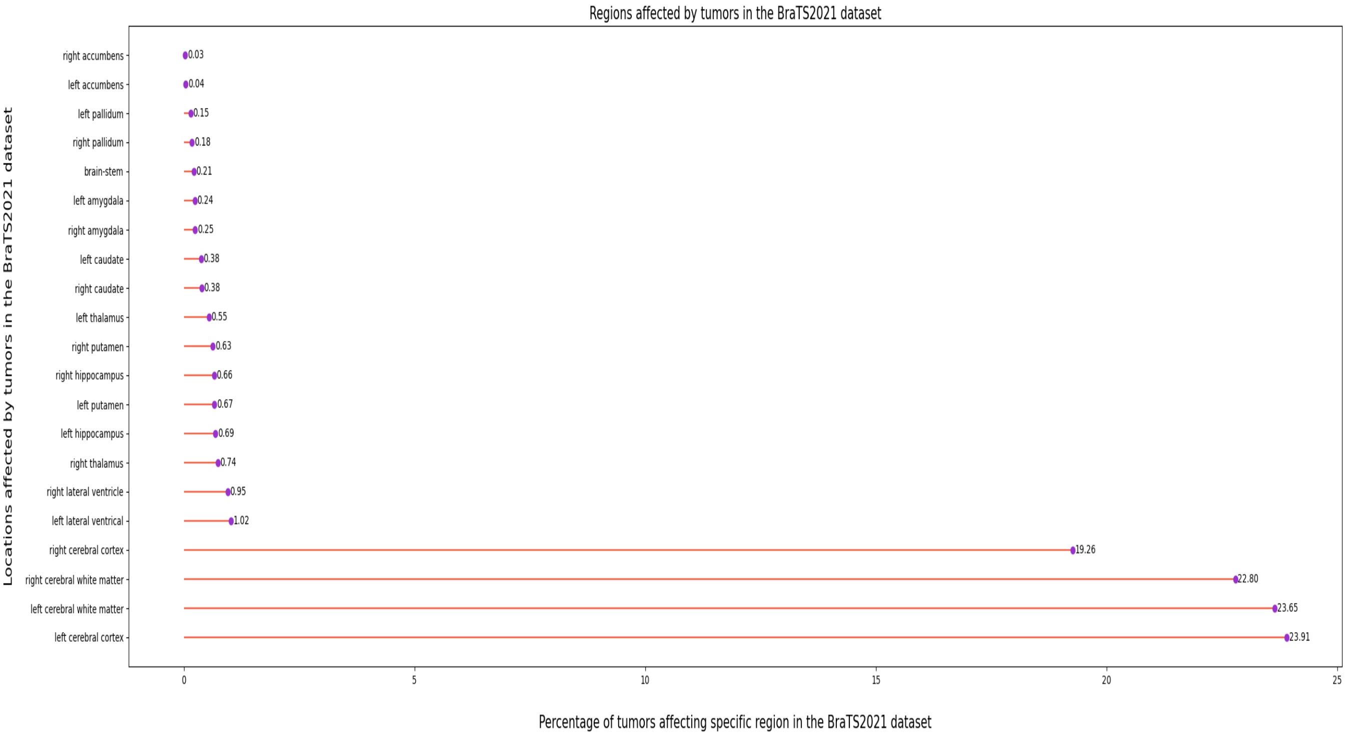


Figure A.10: The subcortical regions, in contrast to non-cortical regions affected by tumors in the BraTS-2021 dataset.

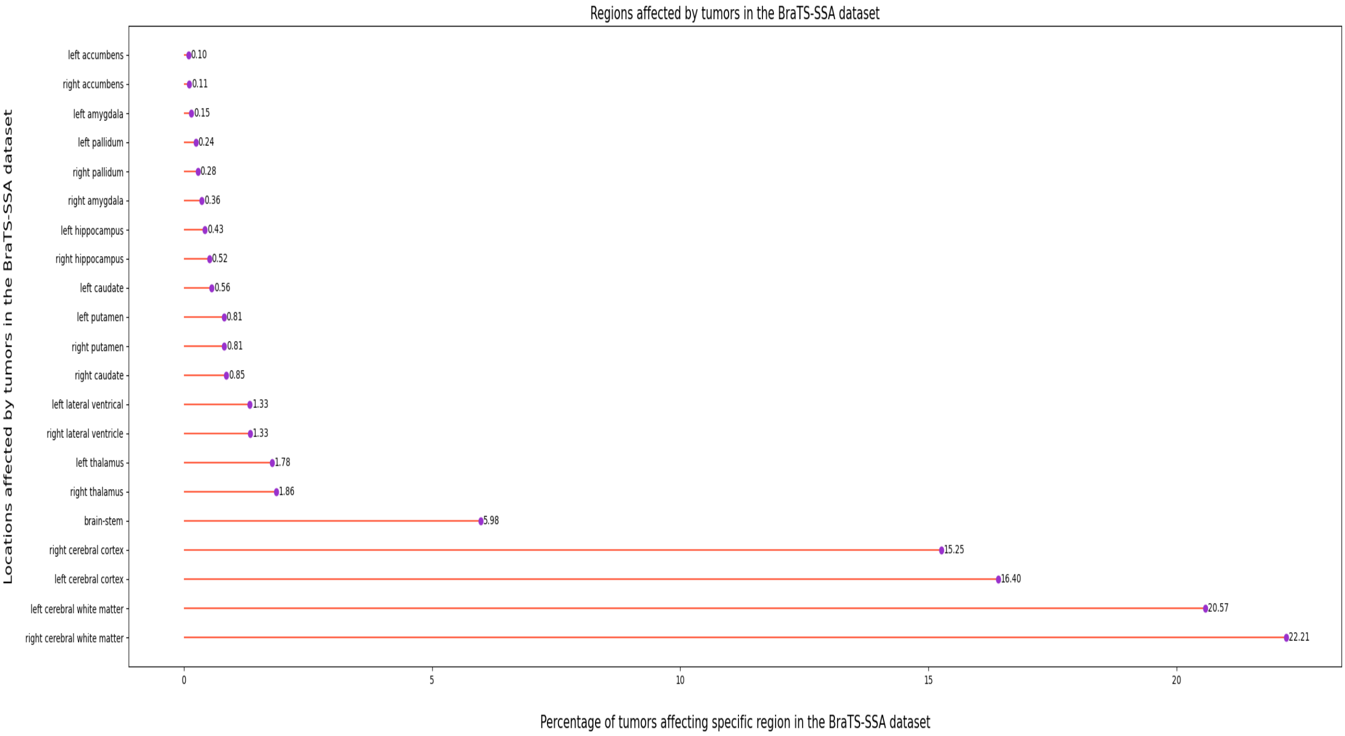


Figure A.11: The subcortical regions, in contrast to non-cortical regions affected by tumors in the BraTS-SSA dataset.

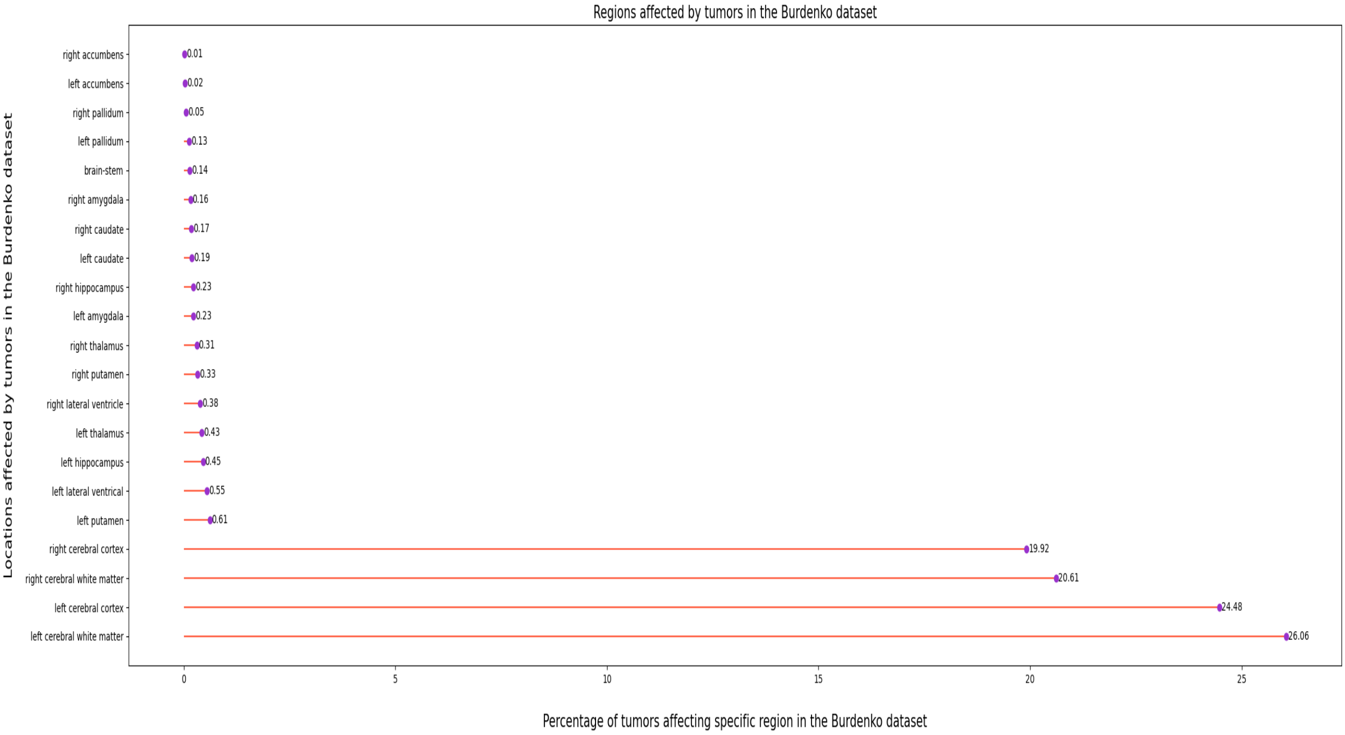


Figure A.12: The subcortical regions, in contrast to non-cortical regions affected by tumors in the Burdenko-GBM dataset.

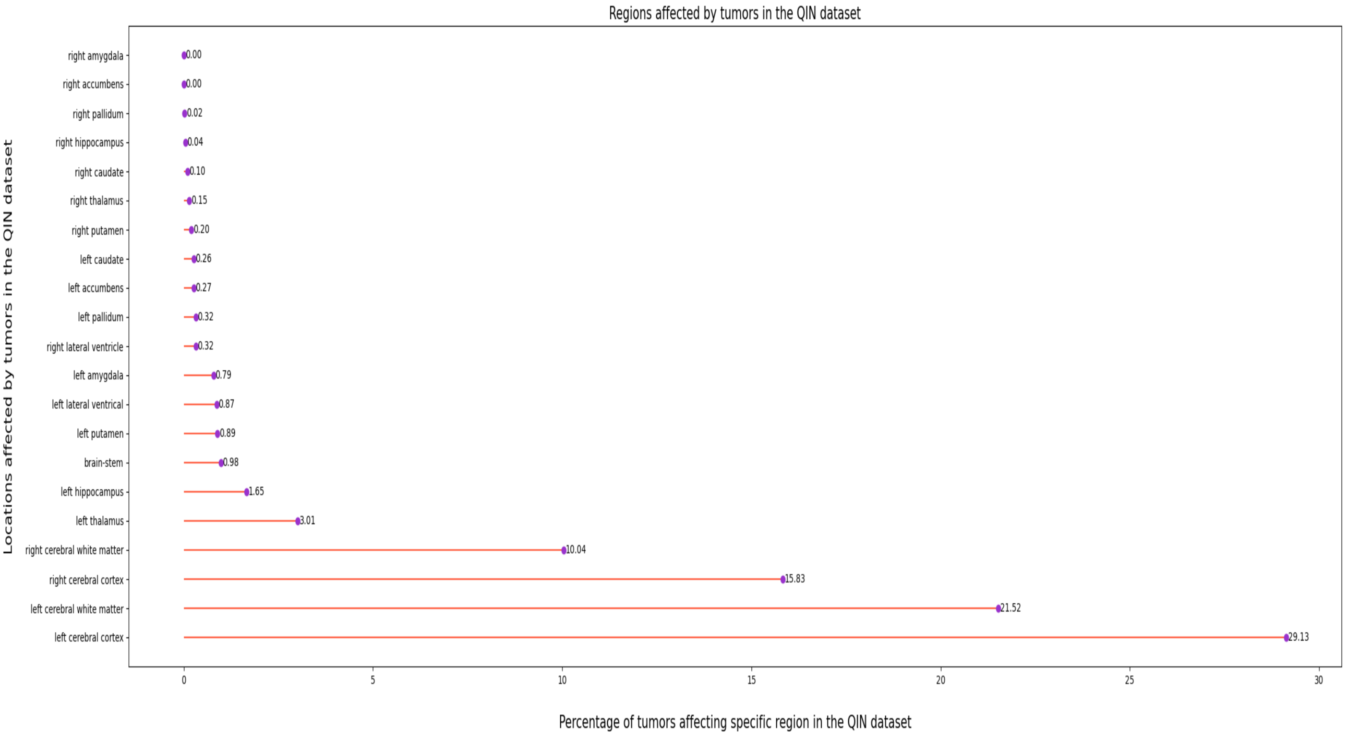


Figure A.13: The subcortical regions, in contrast to non-cortical regions affected by tumors in the QIN dataset.

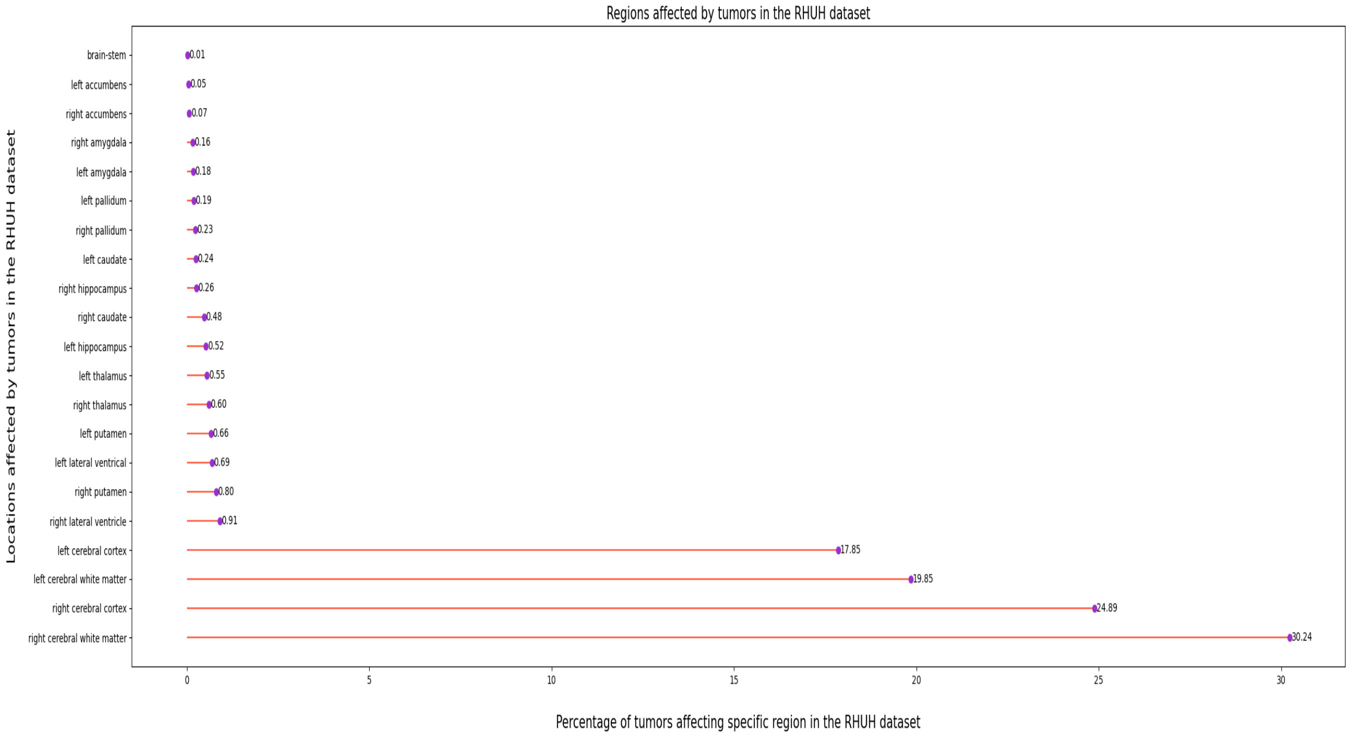


Figure A.14: The subcortical regions, in contrast to non-cortical regions affected by tumors in the RHUH-GBM dataset.

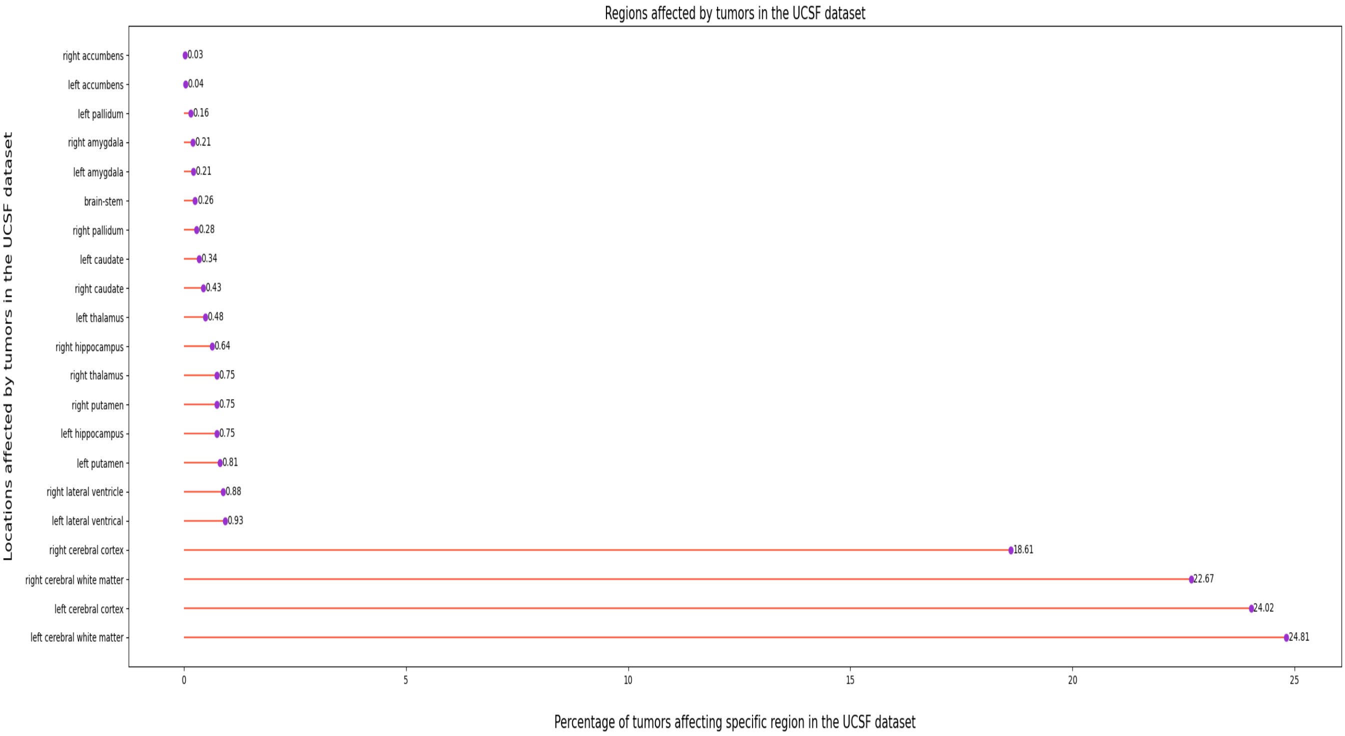


Figure A.15: The subcortical regions, in contrast to non-cortical regions affected by tumors in the UCSF-PDGM dataset.

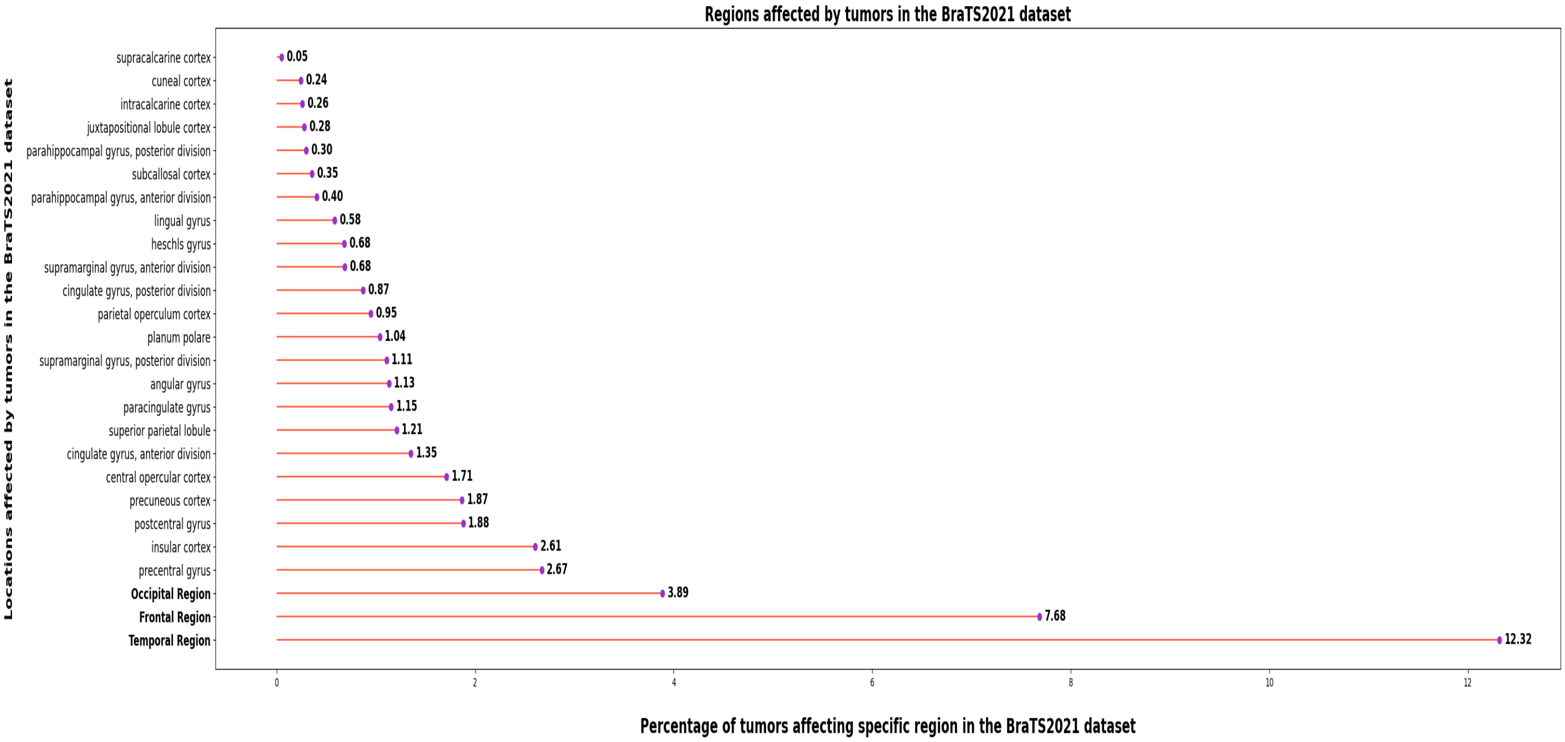


Figure A.16: A detailed view of the cortical regions affected by tumor.

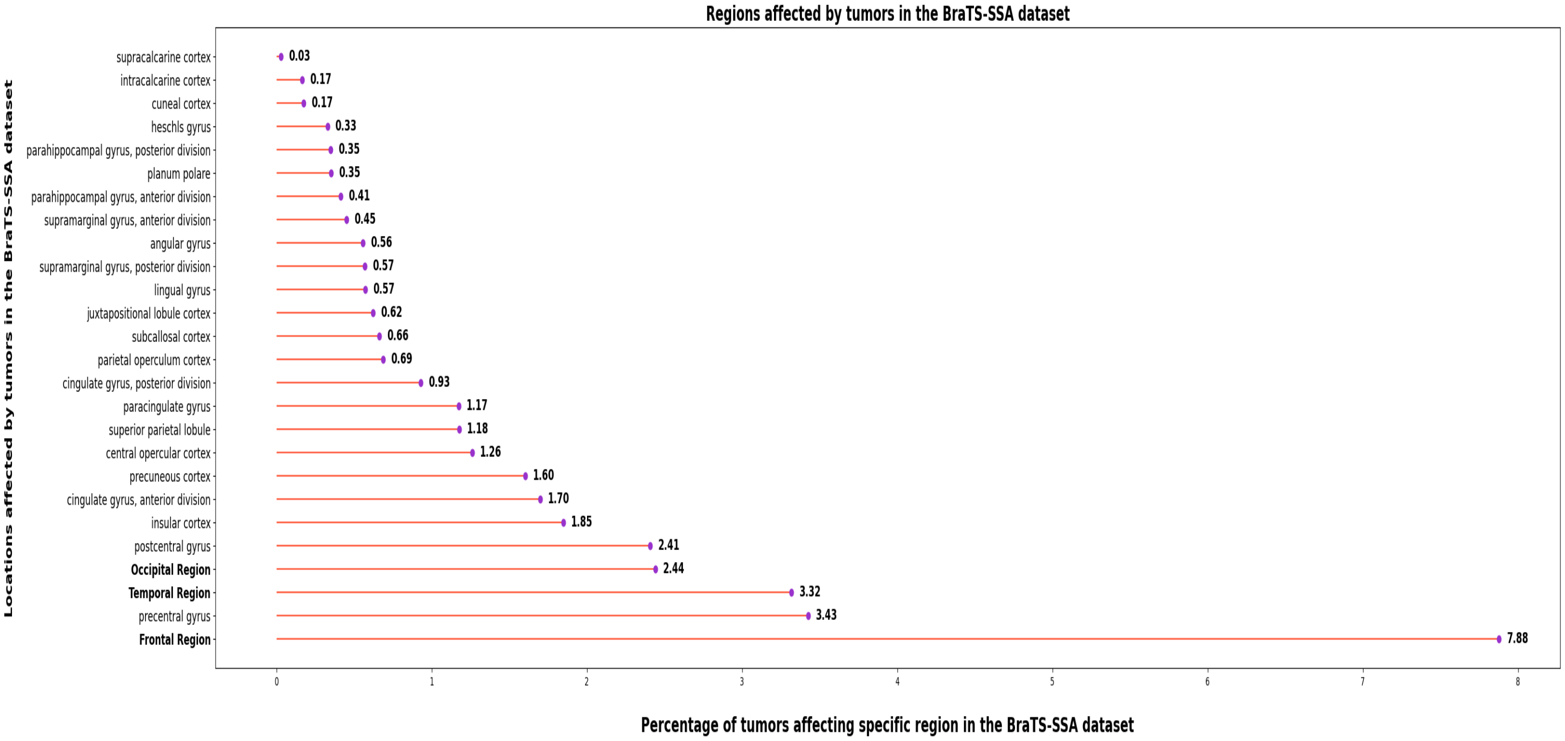


Figure A.17: A detailed view of the cortical regions affected by tumor.

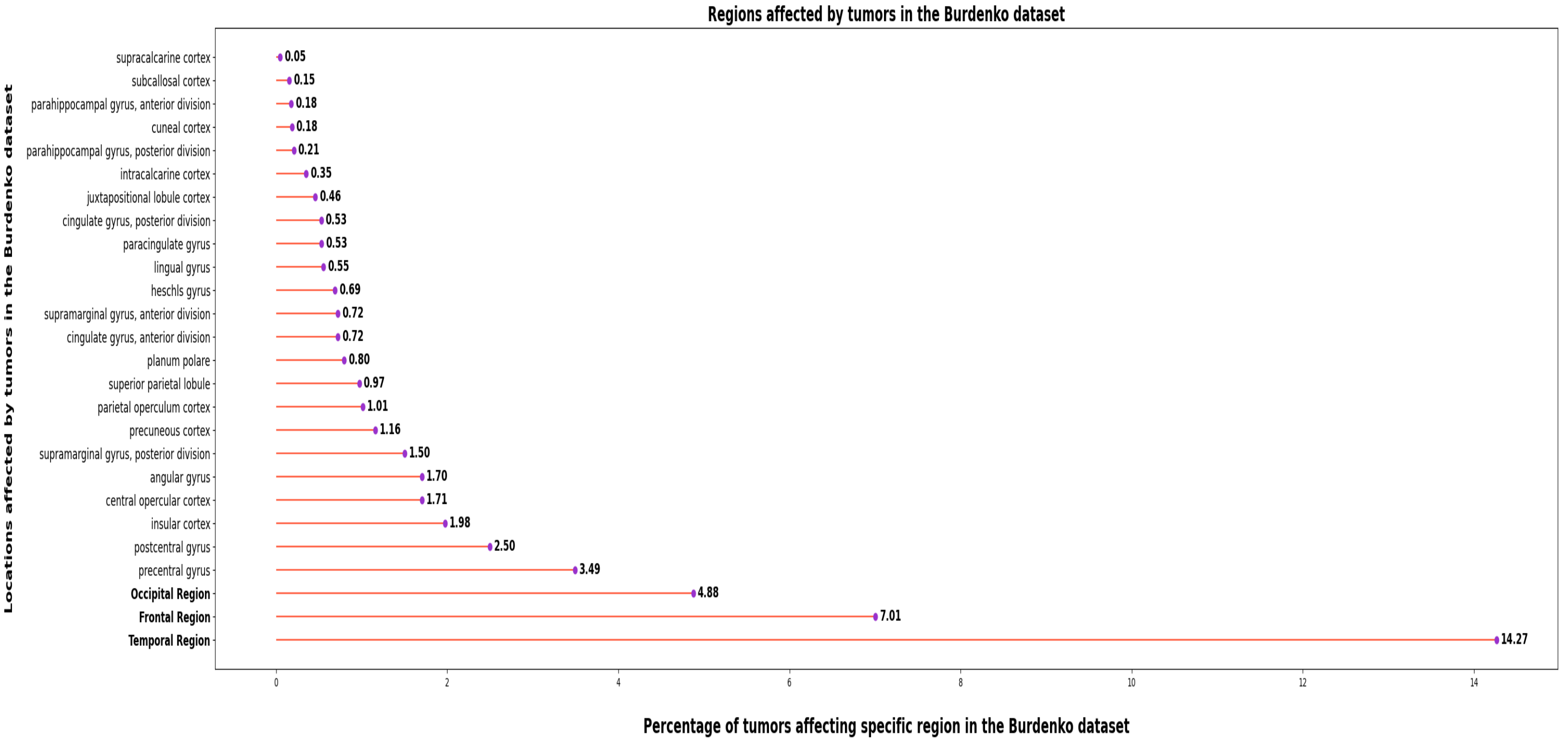


Figure A.18: A detailed view of the cortical regions affected by tumor.

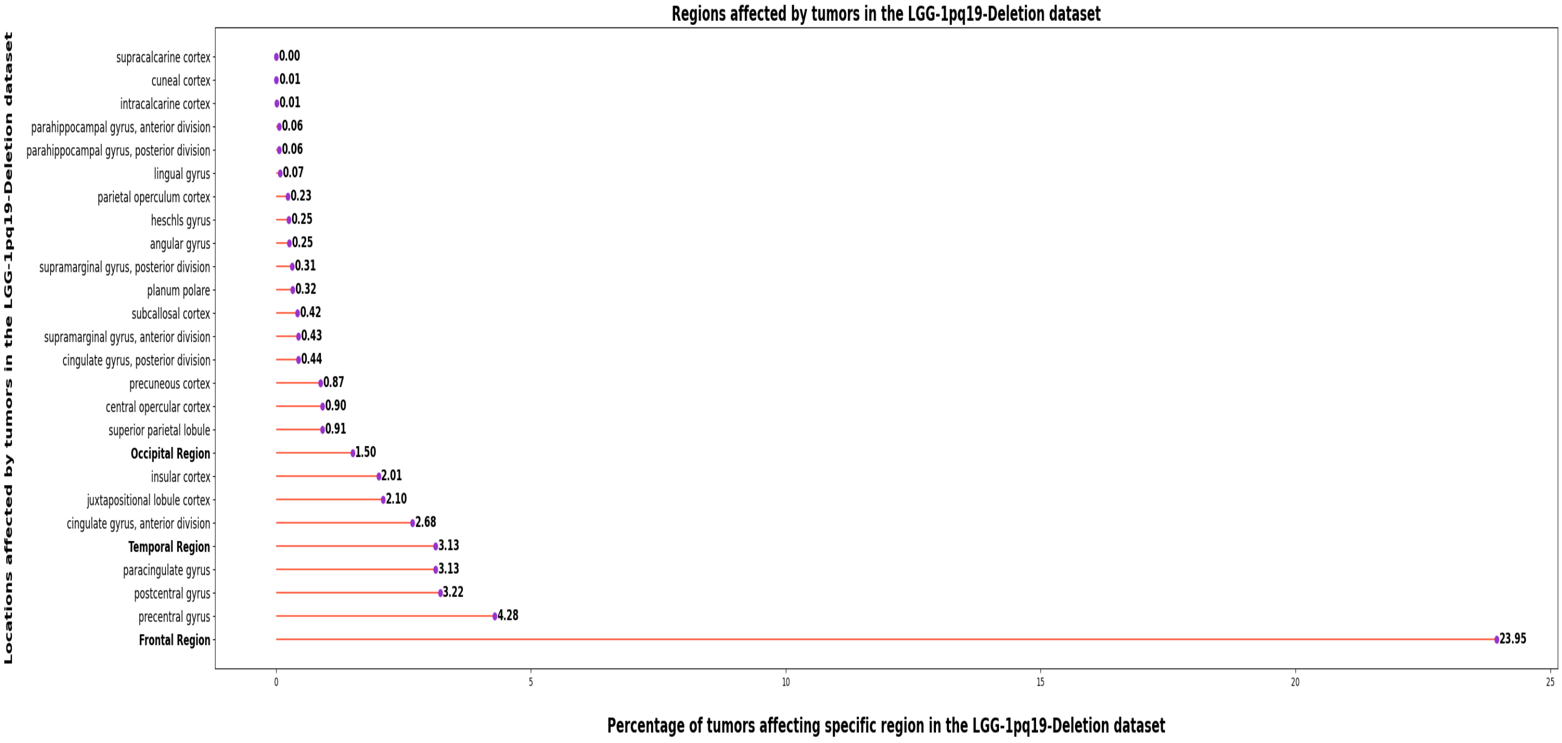


Figure A.19: A detailed view of the cortical regions affected by tumor.

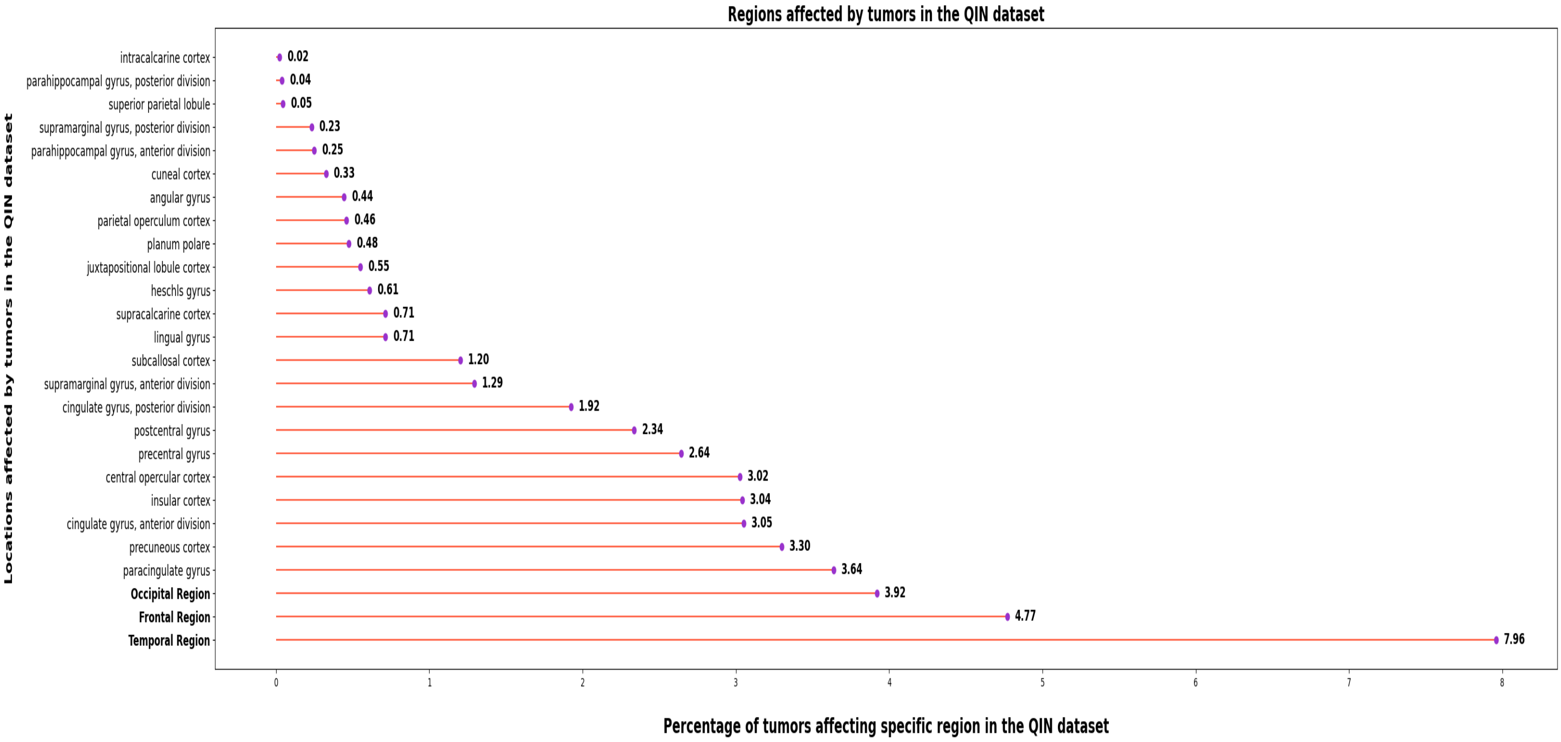


Figure A.20: A detailed view of the cortical regions affected by tumor.

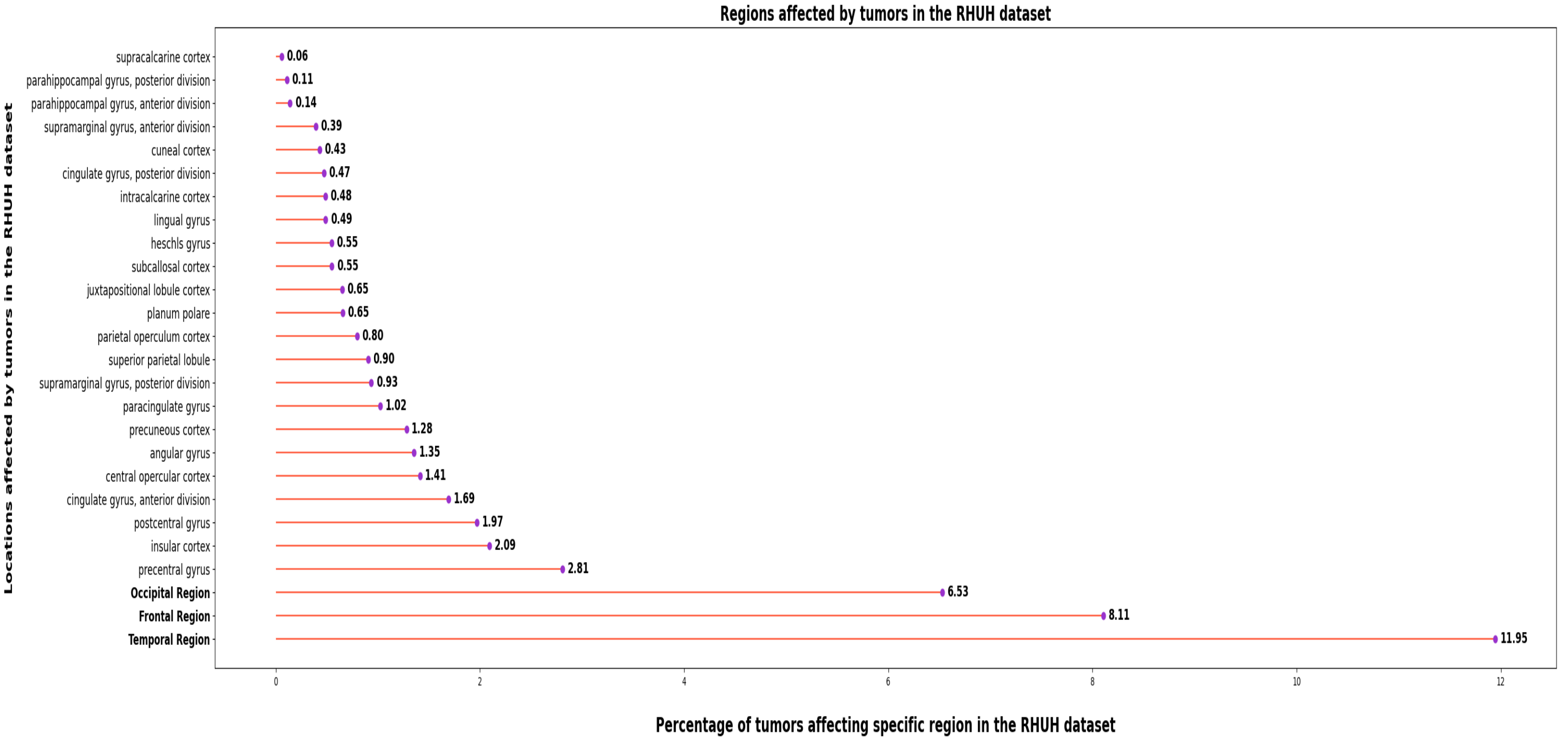


Figure A.21: A detailed view of the cortical regions affected by tumor.

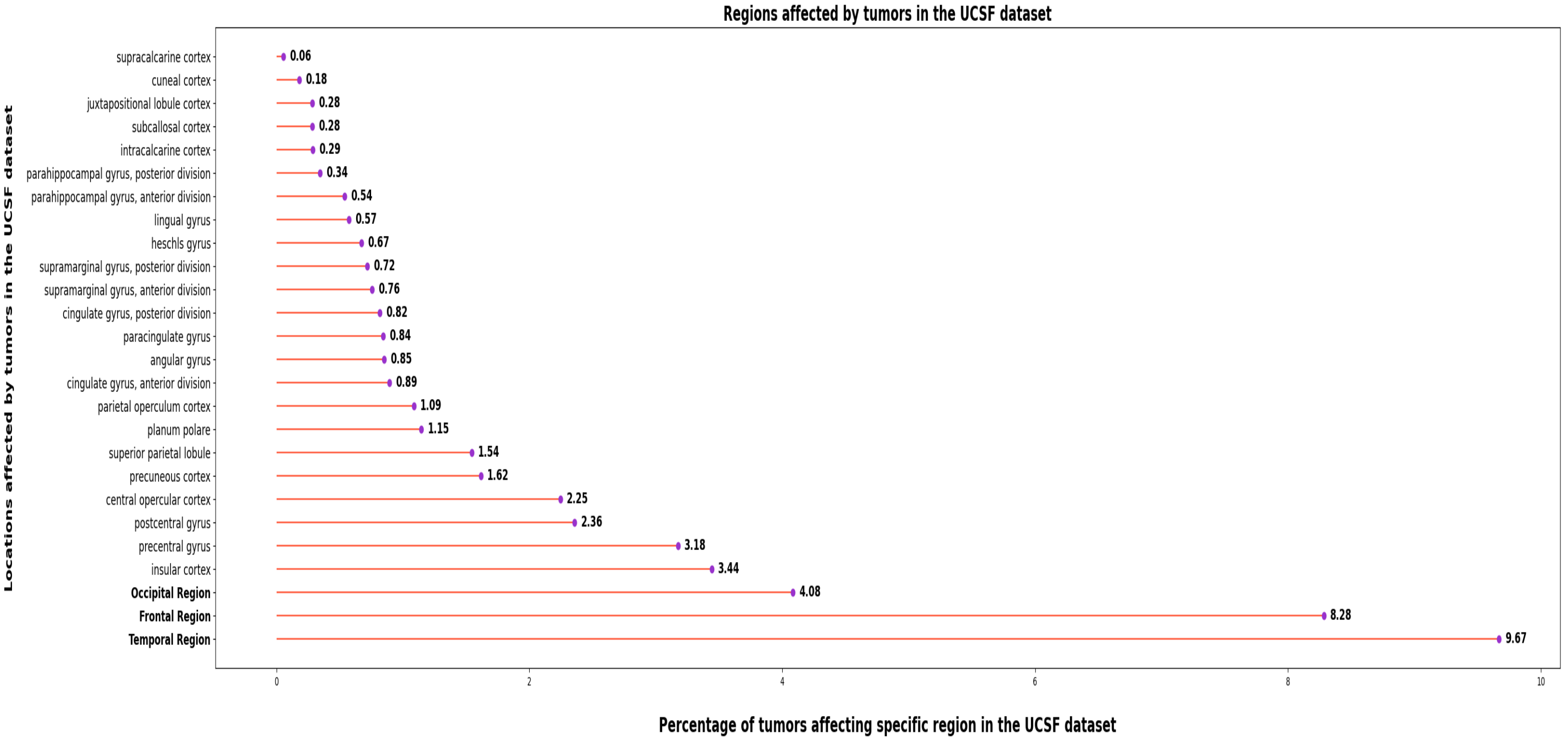


Figure A.22: A detailed view of the cortical regions affected by tumor.

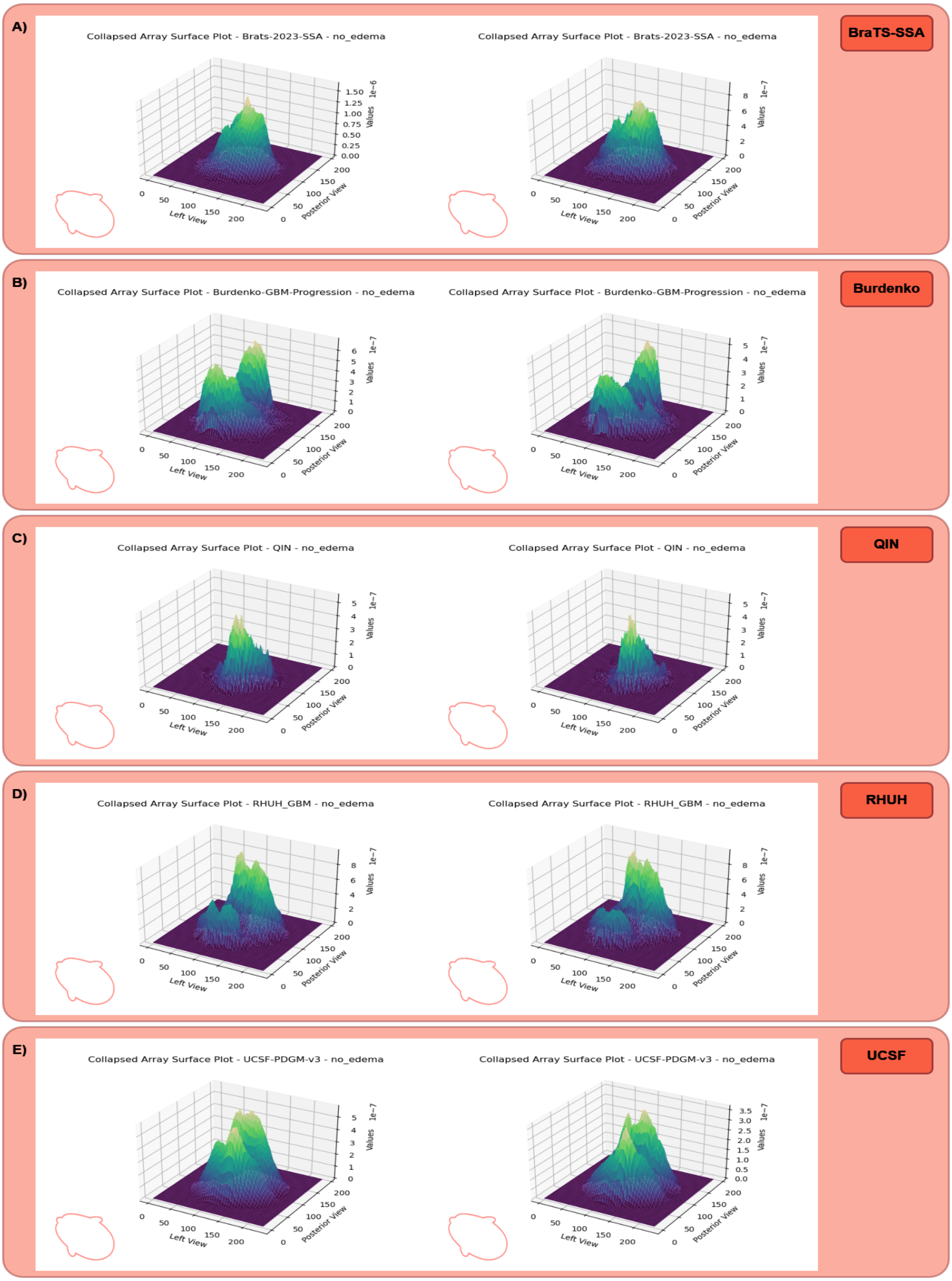


Figure A.23: Supplementary 2D probability distributions. (Left Column) Probability distribution after rigid registration. (Right Column) Probability distribution after deformable registration.

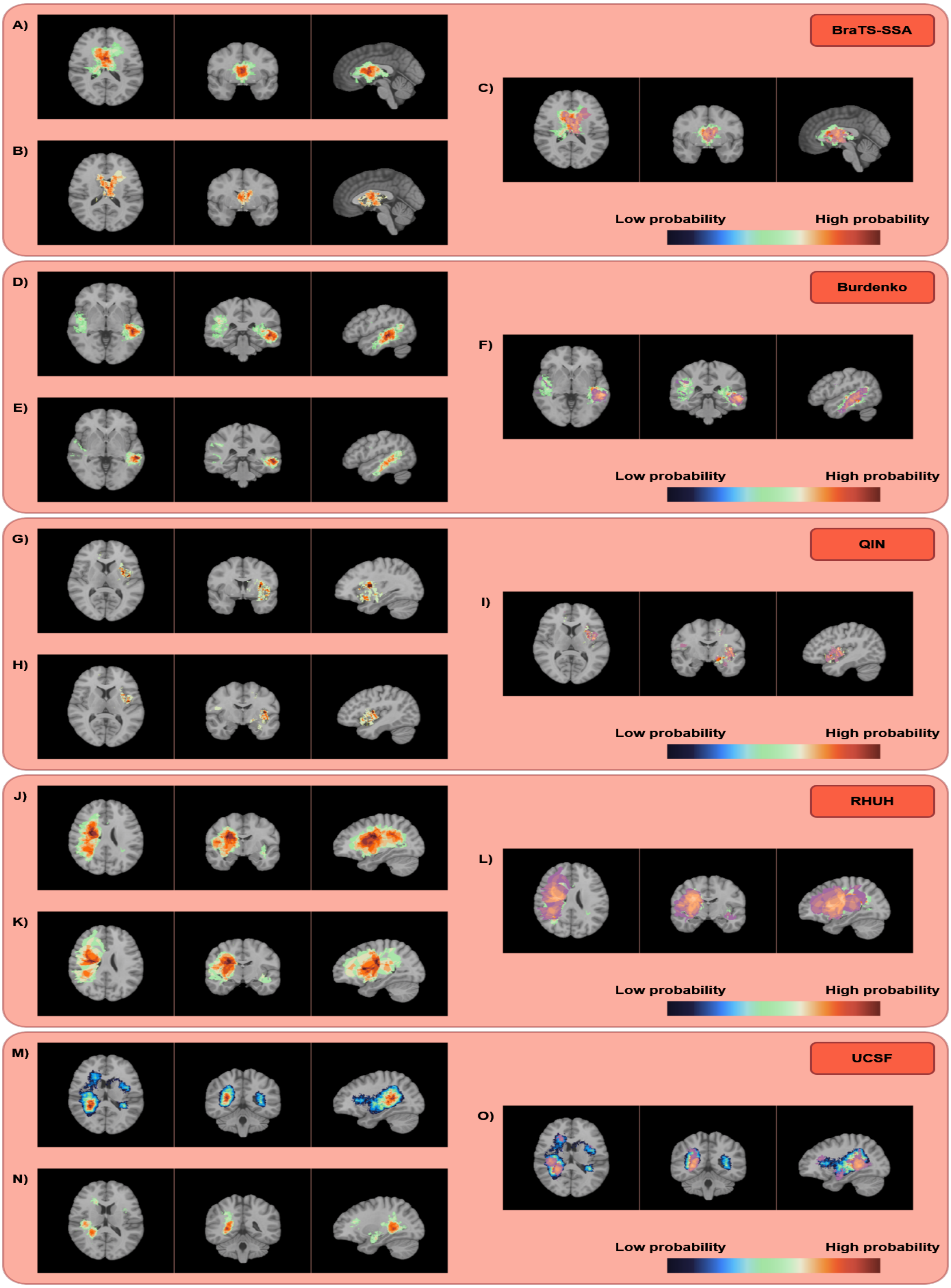


Figure A.24: Supplementary 3D spatial distributions. (First Row) Spatial distribution after rigid registration. (Second Row) Spatial distribution after deformable registration. (Third Row) Comparison of the deformable and rigid registrations. The affine registration kept the jet color scheme, and the deformable registration was given the copper color scheme for contrast.

A.4 Supplementary duplicate recognition figures, tables, and equations

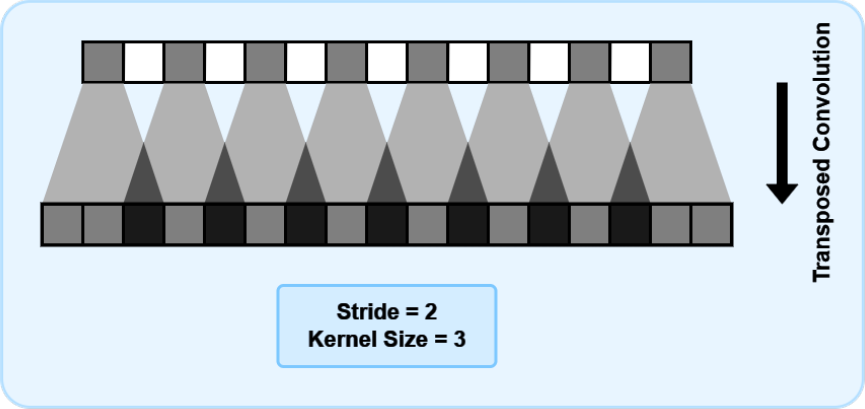


Figure A.25: How a checkerboard artifact is generated from an image (Top) after transposed convolution (Bottom) with a stride of 2 and a kernel size of 3

Loss Function	$Accuracy_{T2a/T2a}$		$Accuracy_{T2}$	
	UNET	**SNN**	**UNET**	**SNN**
Combined Loss	57 %	/	0	/
Photometric Triplet Loss	60 %	/	0	/
Triplet Loss	58 %	59 %	0	0
SSIM Loss	58 %	/	0	/
Photometric Loss	60 %	/	0	/

Table 9: Accuracy of a model trained and evaluated on a contaminated dataset. The dataset contained both T1 images and T2 images. Due to architecture of the training loop, the model was training on identifying T1 images from augmented T1 images, or identifying T2 images from augmented T2 images. Then during testing it was asked to identify T1 images from augmented T1 and augmented T2 images. The results show that the accuracy for identifying T1 from T1 augmented and T2 augmented is around 60%, but this is because the model is correctly identifying the T1s from augmented T1s, and not from T2s. As can be seen from the right column, the model can't identify T1s from T2.

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}, \tag{13}$$

where :

- μ_x is the pixel sample mean of x
- μ_y is the pixel sample mean of y
- $c_1 = (k_1 L)^2$ is a variable to stabilize the division with weak denominator

- L is the dynamic range of the pixel-values
- $k_1 = 0.01$

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}, \tag{14}$$

where :

- σ_x^2 is the variance of x
- σ_y^2 is the variance of y
- $c_2 = (k_2L)^2$ is a variable to stabilize the division with weak denominator
- L is the dynamic range of the pixel-values
- $k_2 = 0.03$

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}, \tag{15}$$

where :

- σ_x^2 is the variance of x
- σ_y^2 is the variance of y
- $c_2 = (k_2L)^2$ is a variable to stabilize the division with weak denominator
- $c_3 = c_2/2$