Biomedical Data Science Capstone Project

Introduction: Medical image segmentation is a critical task in modern healthcare, enabling precise delineation of anatomical structures and pathological regions from imaging modalities such as MRI and CT scans. Accurate segmentation facilitates diagnosis, treatment planning, and longitudinal disease monitoring, particularly in conditions like brain tumors, liver lesions, and lung abnormalities. However, developing robust and generalizable segmentation algorithms remains a significant challenge due to the variability in image quality, scanner types, acquisition protocols, and patient populations across clinical sites. To address these challenges, the Medical Segmentation Decathlon (MSD) was established as a large-scale, standardized benchmark for evaluating medical image segmentation methods across multiple modalities and anatomical regions.[1,2] The dataset encompasses ten diverse tasks spanning a wide range of organs and pathologies. However, this project will only focus on a data analysis of the training dataset of various modalities of magnetic resonance imaging (MRI) brain scans from Task 01: Brain Tumor Segmentation. The dataset includes MRI scans from patients with glioblastoma and low-grade gliomas. Gliomas are a broad class of primary brain tumors that arise from glial cells, the supportive cells of the central nervous system responsible for maintaining homeostasis, forming myelin, and providing structural support for neurons. They account for roughly 30% of all brain and central nervous system tumors and 80% of malignant brain tumors.[3,4] Gliomas are classified by the World Health Organization into grades I–IV based on various features such as cellularity, mitotic activity, necrosis, and vascular proliferation.[3] Glioblastoma represents the most aggressive and lethal form of glioma (grade IV). It is characterized by rapid proliferation, diffuse infiltration into surrounding brain tissue, and extensive vascularization and necrosis. Glioblastoma typically presents with poor prognosis and a median survival of only 12–15 months despite existence of treatments involving surgery, radiation, and chemotherapy. [4]

Data Overview: The Brain dataset focuses on the segmentation of brain tumor subregions from MRI scans. Specifically, it targets three key tumor components: the edema, enhancing tumor, and non-enhancing tumor regions. The dataset includes four MRI sequences commonly used in neuroimaging: native T1-weighted (T1), Gadolinium-enhanced T1-weighted (T1-Gd), native T2-weighted (T2), and T2 Fluid-Attenuated Inversion Recovery (FLAIR), which together provide complementary information for distinguishing between healthy tissue and various tumor substructures. All the images in the data were already preprocessed to be uniform. The complete dataset is divided into 484 training cases and 266 test cases.[2] However, for the purpose of this data analysis, only the 484 training cases were used.

Analysis: For each patient, tumor volume was computed by summing all voxels belonging to non-zero segmentation labels. Across the 484 patients in the training dataset, it was found that the mean total tumor volume was around 103,000 $\pm$ 60,000 voxels, with a median of around 94,000 voxels. The volume of the tumors ranges from around 7,000 voxels up to over 300,000 voxels. (See Figure 1)

	Tumor Volume in Voxels
Mean	103299.617769
STD	59269.356479
Min	7285.000000

Median	94221.500000
Max	318345.000000

Figure 1: Table of descriptive statistics for the volume of the tumors (across all 484 patients in the training dataset).

Across all patient cases, the enhancing tumor region exhibits a distribution most closely matching that of the total tumor volume. Both share a similar shape and central tendency, suggesting that the enhancing region is the major contributor to the overall tumor. In contrast, the peritumoral edema and necrotic core components display right-skewed distributions, indicating that for most patients these regions occupy relatively small volumes, with only a few outlier cases showing much larger extents. (See Figure 2)

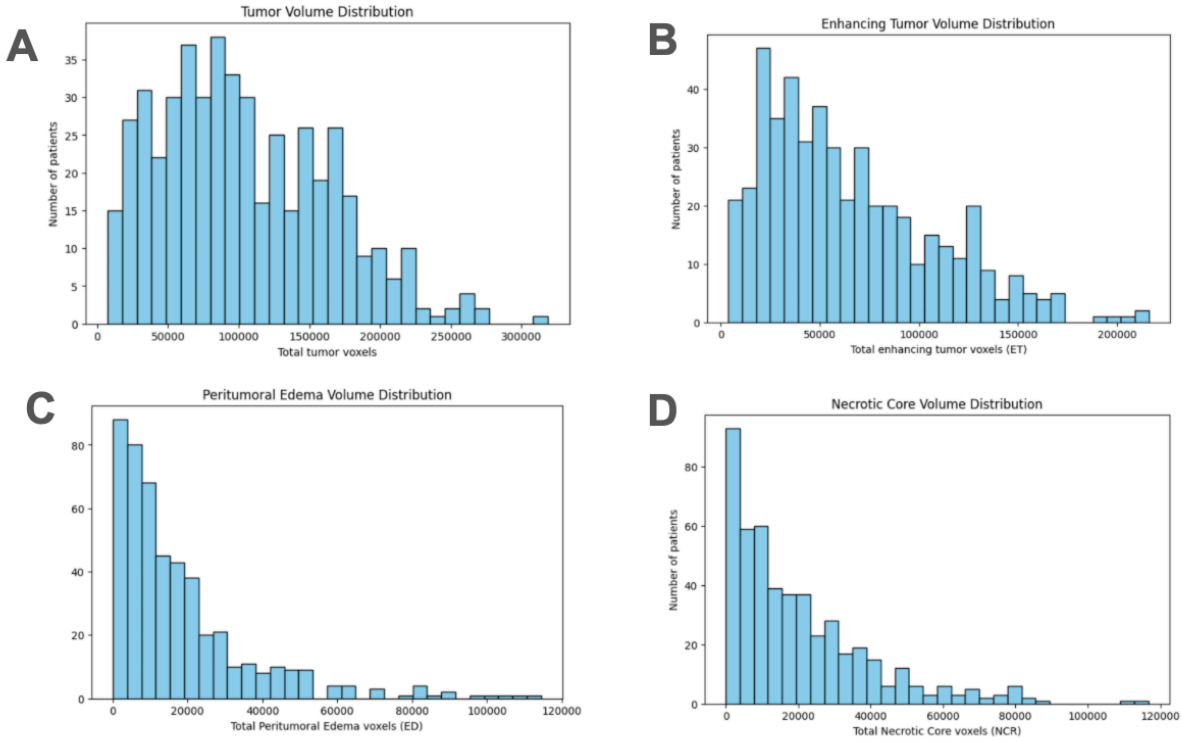

Figure 2: Bar charts of tumor volumes across all patients. [A] total volume, [B] Enhancing, [C] Peritumoral Edema, [D] Necrotic Core

When visualizing tumor volumes across all patients (sorted by total tumor size), the enhancing tumor consistently emerges as the dominant subregion across nearly all subjects. (See Figure 3) This reinforces the observation that enhancing tumors, corresponding to regions of active tumor growth and blood-brain barrier disruption, are the most substantial and consistent component of gliomas in this dataset.[1]

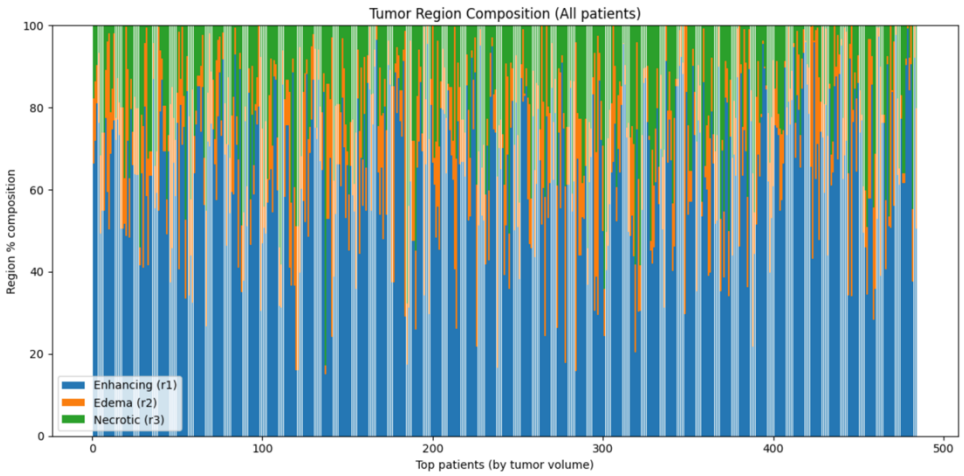

Figure 3: Percent composition of tumors for all patients.

To evaluate hemispheric differences in tumor presentation, left and right hemisphere tumor volumes were computed for each patient. The resulting scatter plot, comparing left versus right tumor volume, revealed a roughly symmetric distribution, suggesting no strong dominance in glioma occurrence across the cohort. (See Figure 4) This finding aligns with existing literature indicating that gliomas can arise in either hemisphere with approximately equal probability, though their clinical symptoms may differ depending on the functional regions affected. [3]

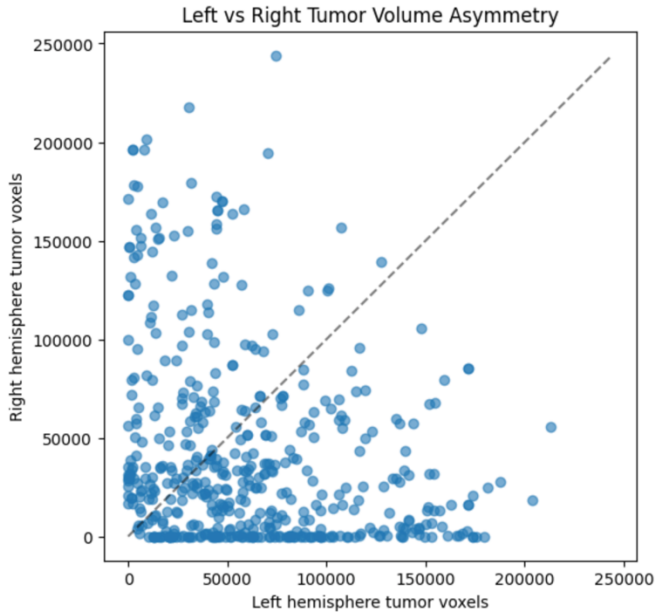

Figure 4: Comparison of right vs left hemisphere tumor volume.

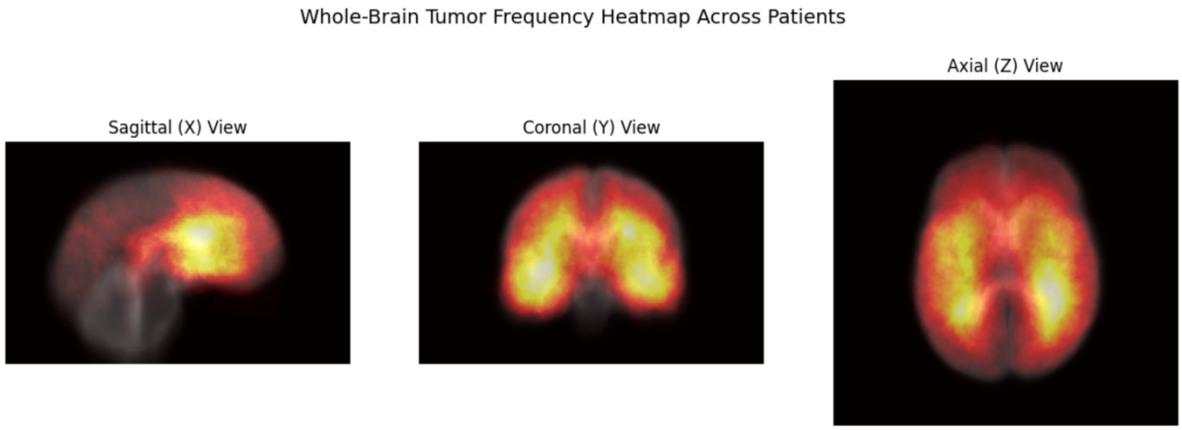

Figure 5: Heat map of common tumor locations across all patients.

To visualize common tumor locations, all segmented masks (the label images for the data) were aligned and averaged voxel-wise across all 484 patients. The resulting heat map reveals the spatial density of tumor occurrence. Yellow regions indicate high overlap, meaning that these are locations where tumors frequently appear. The red regions indicate moderate to low overlap, meaning tumors appear at moderate to low frequencies. (See Figure 5) The heat map shows tumors predominantly centered in frontal and temporal lobes, which is consistent with clinical trends in gliomas. [3,4]

Limitations: There are several limitations to this analysis. The data used for this analysis came from many different sources, resulting in many of the images being obtained under different imaging protocols, with differences in annotation procedures. To ensure consistency for the challenge, pixel level annotations were completed by one individual.[1,2] Including multiple annotators would have enabled estimation of reliability and potentially improved annotation consistency, but this was not possible given the dataset's scale and resource constraints.[1,2] Previous studies have also shown that multiple annotators are often necessary to reduce variability of annotations between raters.[5] The dataset does not include patient demographics, tumor grades, or survival outcomes, which restricts the ability to perform correlation analyses between imaging features and clinical outcome. Additionally, the analysis is limited by the lack of longitudinal data. This prevents evaluation of tumor progression, treatment response, or recurrence over time. Thus, the use of this data is essentially limited to segmentation and visualization.

Future Analysis: A logical future step to this project would be to implement a deep learning-based image segmentation task, similar to the mouse nuclei segmentation project completed in Biomedical Data Science Lab 3. In that assignment, segmentation was performed using TensorFlow to classify and segment nuclei regions from microscope images. A comparable task could be applied here to segment the three tumor subregions (enhancing, peritumoral edema, and necrotic core) from the provided multi-modal MRI data. This would also involve constructing and training a U-Net model using the provided MRI images as input tensors and the provided labeled images as targets. This step would automate the segmentation process, which is an area of interest for many clinicians.

References

1. Antonelli, M., Reinke, A., Bakas, S., et al. (2022). The Medical Segmentation Decathlon. *Nature Communications*, *13*, Article 4128. https://doi.org/10.1038/s41467-022-30695-9
2. Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., Bilic, P., Christ, P. F., Do, R. K. G., Gollub, M., Golia-Pernicka, J., Heckers, S. H., Jarnagin, W. R., ... Cardoso, M. J. (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint* arXiv:1902.09063. https://arxiv.org/abs/1902.09063
3. Cini, N. T., Pennisi, M., Genc, S., Spandidos, D. A., Falzone, L., Mitsias, P. D., Tsatsakis, A., & Taghizadehghalehjoughi, A. (2024). Glioma lateralization: Focus on the anatomical localization and the distribution of molecular alterations (Review). *Oncology reports*, *52*(4), 139. https://doi.org/10.3892/or.2024.8798
4. Larjavaara, S., Mäntylä, R., Salminen, T., Haapasalo, H., Raitanen, J., Jääskeläinen, J., & Auvinen, A. (2007). Incidence of gliomas by anatomic location. *Neuro-oncology*, *9*(3), 319–325. https://doi.org/10.1215/15228517-2007-016
5. Joskowicz, L., Cohen, D., Caplan, N. *et al.* Inter-observer variability of manual contour delineation of structures in CT. *Eur Radiol* **29**, 1391–1399 (2019). https://doi.org/10.1007/s00330-018-5695-5