Helen Tian
MACS 23002
Research Project Proposal

- *What is the research question? What motivates you to do this research? Provide at least one related paper.*

  How can we accurately classify MRI images of brain tumors into categories of glioma, meningioma, no tumor, and pituitary? How can we estimate the tumors' volume from multimodal MRI scans?

  1. The first objective is to develop a robust classification model that can accurately classify MRI images of the brain into the specified categories of glioma, meningioma, no tumor, and pituitary. We are using a dataset that has already been categorized and labeled with images and their respective tumors (detailed more below).
  2. The second objective is to investigate the feasibility of using symmetry-based methods to estimate tumor volume from the MRI images. This could involve analyzing the symmetry of tumor structures within the images and leveraging this information to infer the overall volume of the tumor from a 2-dimensional scan.

  Related paper: Automatic Brain Tumor Detection and Volume Estimation in Multimodal MRI Scans via a Symmetry Analysis (https://www.mdpi.com/2073-8994/15/8/1586)

  - This study introduces an automated medical decision support system for brain tumor detection, segmentation, and volume estimation from MRI images. It utilizes unsupervised approaches based on histogram and symmetry analyses, achieving high accuracy rates without the need for extensive labeled training data, which could significantly enhance surgical planning and treatment outcomes for brain tumor patients. It gave me the idea for my research project, especially the volume estimation part; I wanted to calculate the potential of using such methods to estimate the tumor volume with a different dataset (from the one they used).
  - I also gained inspiration for a lot of how to pixelate and find a lot of the features that I talk about more below through the following dataset and notebook: www.kaggle.com/balakrishcodes/brain-2d-mri-mask. I hope to use some of what they show in their notebook in my own project.

  As someone currently thinking of pre-med, I thought this was a really interesting project to combine my interests of CS (specifically, machine learning) and medicine. Machine learning has been of rising technological use within the medical field, and I wanted to explore such techniques/datasets through this project. Additionally, I'll be working in a neurobiology lab this summer, so it's nice to get ahead.

- *How to formulate your research question to a machine learning task?*
  - *Is it classification (binary / multi-class) or regression?*

  Both of my tasks (tumor classification and volume estimation) both fall under supervised learning.

  1. The first objective involves a multi-class classification task, where each MRI image is assigned to one of the predefined classes — glioma, meningioma, no tumor, and pituitary.
  2. Since tumor volume is a continuous variable, the task of estimating tumor volume falls under regression. In this case, the input to the regression model would likely be points of tumor pixelation extracted from the MRI images, and the output would be the estimated volume of the tumor.

- *Do you plan to use existing dataset or collect your own?*
  - *If it is an existing dataset, please provide the link and indicate how do you plan to process it using EDA techniques? Does the dataset match with the research question?*

Yes, these datasets match my research question. Here is my dataset: www.kaggle.com/brain-tumor-mri-dataset; this dataset is a combination of the following three datasets: figshare, SARTAJ dataset, Kaggle Br35H. I have outlined the steps of my EDA process below:

1. **Visualization of Class Distribution**
   Plot a bar chart or pie chart to visualize the distribution of classes within the dataset. This will help you understand the balance or imbalance of classes and whether any class is significantly underrepresented.
2. **Sample Image Visualization**
   Display sample MRI images from each class to visually inspect the characteristics and variability of images within each category. I'll use libraries like Matplotlib or Seaborn to visualize the images.
3. **Feature Extraction**
   I will be performing preliminary feature extraction and exploration. This may involve computing basic statistics (e.g., mean intensity, variance) and texture features (e.g., GLCM features) from the images and visualizing their distributions across different classes.
4. **Correlation Analysis**
   Explore correlations between different features extracted from the images and the corresponding tumor categories. This can help identify which features are most informative for tumor classification and volume estimation.
5. **Quality Control**
   Check for any data preprocessing or cleaning steps needed, such as handling missing values, standardizing image sizes, or removing artifacts. Ensure that the dataset is clean and suitable for further analysis. This also includes dimensionality checks/reduction; because our dataset is a compilation of 3 different datasets, I must pay attention to the size of the images in this dataset. I will resize the images to the desired size after pre-processing and removing the extra margins.

- *About the dataset*
  - *What is the data size?*
  - *Do you have labels for the data?*
  - *What's the feature space? How many features? what type of features (E.g., 10 features in total, 8 numerical, 2 categorical)*

  The dataset I'm using contains 7023 images of human brain MRI images which are classified into 4 classes: glioma, meningioma, no tumor, and pituitary. Currently, each MRI image is only represented by 1 categorical attribute (e.g. tumor type), but from my second objective, I think I will be able to dissent the images into a few more numerical features (e.g. geometric properties (e.g. area, perimeter, centroid of tumor), pixel intensity values, texture aka gray-level occur matrix features, etc.)

- *Highlight one exciting aspect of the dataset.*

  One really exciting aspect of this dataset is engaging in the process of feature extraction myself. It was difficult to build a large enough and comprehensive dataset like this one; it provides a lot of hands-on experience to explore different image processing techniques and algorithms.

- *Who is your peer? List at least one useful feedback from your peer.*

  The two classmates I talked with were Jun and Shikang. We were all still planning our research questions and looking into different datasets. However, we were able to clarify that we can use a dataset of fewer people where each scan or image is considered a sample, rather than one person with multiple scans being considered one sample. This was really helpful because for medical research data like mine, it's difficult to find a dataset of at least 5000 annotated and cleaned samples (aka people).

- *Any additional information you would like to share?*

  Because we don't have actual medical opinions and multiple slices of our scans, our volume estimation won't really be as accurate as I'd like. It will also be mainly implementing the equations that the linked related paper delves into; however, I'd still like to classify it as regression as I will be implementing the equations (as "feature extraction") on the training dataset and do the regular regression modeling with the test.