MAIS 202 - Deliverable 1 & 2

Eya Ibrahim

1. Dataset

https://www.kaggle.com/datasets/praneet0327/brain-tumor-dataset

I chose the Brain Tumor MRI Scan Dataset by Praneet0327 because it has an extensive range of datapoints, as well as clear categorization into positive (tumor present) and negative (no tumor) classes, which aligns well with my binary classification goals.

This dataset includes over 5,000 MRI images, providing a robust representation of both healthy and tumorous brain structures. The detailed MRI scans enhance the dataset's suitability for ML applications aimed at brain tumor detection and healthcare research. Additionally, its usability rating and manageable file size make it ideal for efficient processing and accurate model training.

2. Methodology

a. Data Preprocessing

The primary data format consists of MRI scans, which will provide visual images that are needed to train my model. Data preprocessing includes cleaning up the data. I would first ensure the images are in the same format, standardize the size, remove corrupted files, and make sure labels are accurate and consistent.

b. Machine Learning Model

The goal is to classify MRI scans into two categories: presence or absence of brain tumors. I think a CNN would be the best fit, a deep learning model that is a good choice for image recognition tasks, given its ability to capture spatial and positional information within images.

CNNs are highly effective for medical image classification due to their hierarchical feature extraction, allowing them to detect shapes, textures, and patterns (such as those found in tumors). The pro is their high accuracy and performance for image classification tasks and the con would be that it needs more computational power, which may require a GPU.

c. Evaluation Metric

<u>Confusion Matrix:</u> For information on true positive, true negative, false positive, and false negative rates.

Accuracy, Precision, and Recall: Accuracy will give an overall correctness measure, while precision and recall are essential to assess the model's performance in detecting tumor-positive cases, minimizing false negatives.

The baseline goal is to achieve an accuracy of at least 85%, with a recall of over 90%, given the high importance of correctly identifying all positive cases.

3. Application

The model will be uploaded as a web application, where the idea is that a radiologist (or perhaps even patient) can directly upload an MRI scan for analysis, cutting down on time and costs.

- User Input: Users will have an option to upload MRI scans (JPEG or PNG format). The web interface will guide users through the upload process and image requirements.
- User Output: After preprocessing and processing, the model will display a diagnosis of either "Tumor Detected" or "No Tumor Detected," along with a probability score (e.g., 95% confidence). This output will be presented in a clean, easy-to-understand interface, so that it is able to be used by a diverse range of users.

This interface will enable healthcare professionals and researchers to quickly assess MRI scans, potentially accelerating preliminary diagnoses and patient triage.