Project Scoping
Brain Tumor Detection

Team Members:
- Aadarsh Siddha
- Akshita Singh
- Vivek Radadiya
- Praneith Ranganath
- Shaun Kirthan
- Yashasvi Sharma

I. Introduction

Brain tumors represent a significant health challenge, ranking as the 10th leading cause of death for both men and women. In 2023, approximately 24,810 adults in the United States were diagnosed with brain tumors, and this number is projected to increase to 30,000 per year. Globally, the incidence of primary brain or spinal cord tumors reached 308,102 in 2020. Given the rising prevalence, early and accurate detection is crucial for improving patient outcomes.

The complexity of brain tumors, with significant variability in their size, location, and nature, poses challenges in fully understanding and accurately diagnosing them. MRI analysis, essential for diagnosis, typically requires the expertise of professional neurosurgeons. However, in developing countries, there is often a shortage of skilled doctors, and the lack of specialized knowledge about tumors can lead to delays and inaccuracies in generating MRI reports.

To address these challenges, this project leverages advancements in machine learning (ML) to develop a comprehensive end-to-end ML pipeline. This automated system will assist doctors in diagnosing brain tumors more effectively, especially in regions with limited access to skilled medical professionals. By deploying this system on the cloud, we aim to provide scalable, reliable, and timely diagnostic support, ultimately improving patient outcomes and reducing the burden on healthcare systems.

II. Dataset Information

1. Dataset Introduction

The dataset for this project combines three sources: figshare, SARTAJ, and Br35H. It consists of MRI images of human brains, categorized into four classes: glioma, meningioma, no tumor, and pituitary. The dataset's purpose is to facilitate the development of a robust model for detecting and classifying brain tumors, supporting medical professionals in making accurate diagnoses.

2. Data Card

Dataset Name: Brain Tumor MRI Images
Size: 7023 images
Format: JPEG
Data Types: MRI images of human brains
Classes: Glioma, Meningioma, No Tumor, Pituitary
Source Datasets: figshare, SARTAJ, Br35H

3. Data Sources

fig share Dataset:
https://figshare.com/articles/dataset/brain_tumor_dataset/1512427
SARTAJ Dataset:
https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri
Br35H Dataset:
https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection?select=no

4. Data Rights and Privacy

All data used in this project are sourced from publicly available datasets with proper usage permissions. We ensure compliance with data protection regulations.

III. Data Planning and Splits

The data planning for this project involves the following sub-steps:

Data Acquisition:

- ➢ Identify and acquire relevant datasets from multiple sources.
- ➢ Verify the quality and integrity of the datasets.

Data Preprocessing:

- ➢ Convert images to a standard format and size.
- ➢ Normalize pixel values.
- ➢ Remove any corrupted or irrelevant data.

Data Augmentation:

- ➢ Apply transformations such as rotation, flipping, and scaling to increase the dataset size.
- ➢ Use library TensorFlow for augmentation.

Data Splitting:

- ➢ We will split the dataset into training, validation, and test sets
- ➢ In a typical 70-20-10 ratio.
- ➢ Stratified sampling will be used to ensureeach class is adequately represented across all splits.

Data Version Control:

- ➢ Use DVC (Data Version Control) to track changes to the dataset.
- ➢ Store data and metadata in a version-controlled environment.

IV. GitHub Repository

1.1 **URL**: GitHub - Omii2899/Brain-Tumor-Classification
1.2 **Folder structure**:
1.2.1 **README.md**: Includes essential project information, installation instructions, and usage guidelines.
1.2.2 **requirements.txt**: Information about the packages and versions required for the project
1.2.3 **src**: All preprocessing and training code will be placed in src folder
1.2.4 **test**: all test files will be placed in this folder
1.2.5 .gitignore
1.2.6 **.github/workflows:** Includes workflow yaml files to automate build, test and deployment pipeline

V. Project Scope

The objective of this project is to design, develop, and deploy a robust ML pipeline capable of detecting brain tumors from medical imaging data. The pipeline will be user-friendly, and scalable, providing valuable insights to healthcare professionals.

The key components of the project include:

- ➢ Data Collection and Preprocessing: Aggregate a large dataset of brain MRI scans from various sources, ensuring data diversity and quality. Preprocess the images to standardize formats, enhance image quality, and annotate tumors accurately.
- ➢ Model Development: Utilize state-of-the-art deep learning techniques, such as Convolutional Neural Networks (CNNs), to build and train a model for tumor detection. Experiment with different architectures and hyperparameters to optimize performance.
- ➢ Model Evaluation and Validation: Implement rigorous evaluation metrics, including accuracy, precision, recall, and F1 score, to assess model performance. Conduct cross-validation and utilize external validation datasets to ensure generalizability.
- ➢ Deployment: Develop an API and a user-friendly interface for healthcare professionals to upload MRI scans and receive diagnostic results. Ensure the system is secure, scalable, and complies with healthcare regulations.
- ➢ Monitoring and Maintenance: Establish monitoring tools to track model performance in real-time and detect any degradation in accuracy. Implement mechanisms for continuous learning and model updates based on new data.

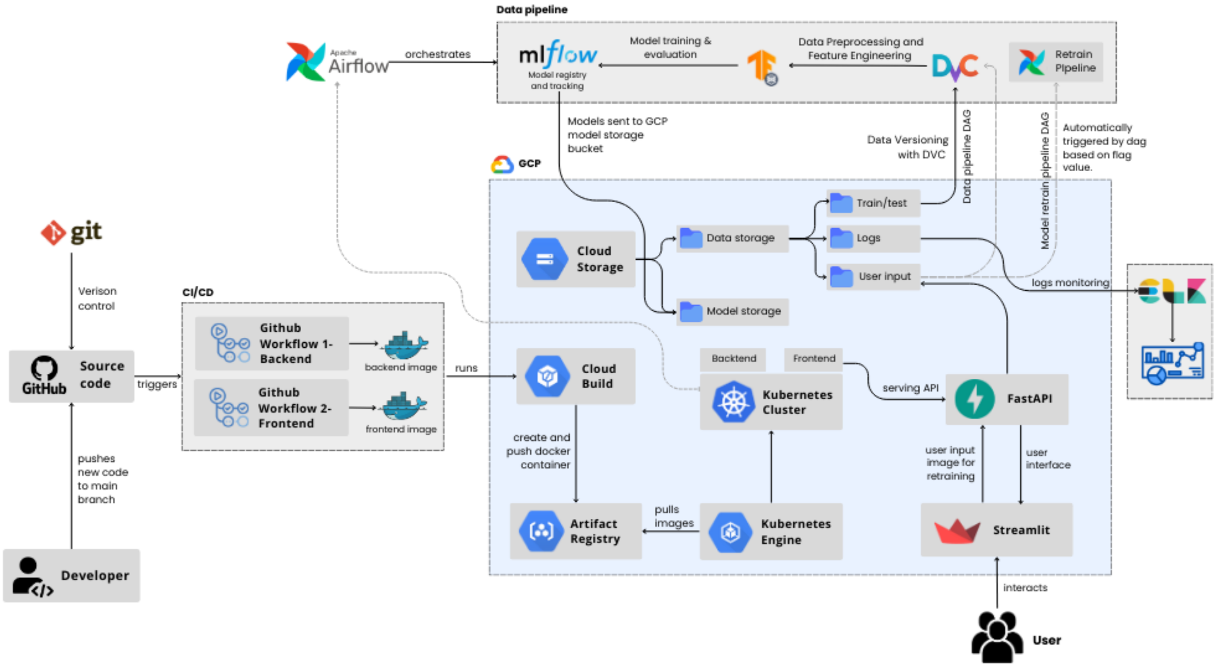
Challenges-

- ➢ Data Quality and Availability: Obtaining diverse, high-quality data while navigating privacy regulations is challenging.
- ➢ Model Generalization: Ensuring consistent performance across different populations and MRI machines is crucial.
- ➢ Computational Resources: Optimizing resource usage for training deep learning models on large datasets is challenging.
- ➢ Regulatory Compliance: Adhering to healthcare regulations like HIPAA for patient data privacy and system reliability is essential.

Proposed Solutions

- ➢ Implement a CNN-based multi-task classification model to perform detection, classification by type within a single framework.
- ➢ Use advanced data augmentation and preprocessing techniques to enhance model performance.
- ➢ Integrate the system into a user-friendly interface for clinical use.

VI. Current Approach Flow Chart and Bottleneck Detection



➢ Limited Access to Annotated Data: High-quality, annotated medical imaging data is scarce, making it difficult to train and validate models effectively.

➢ Variability in Imaging Techniques: Differences in MRI machines and imaging protocols across institutions can lead to variability in the data, affecting model performance.

➢ Interpretability of Models: Deep learning models, especially CNNs, are often seen as black boxes. Improving model interpretability to gain the trust of healthcare professionals is a significant bottleneck.

➢ Scalability and Real-Time Processing: Deploying the model in a clinical setting requires real-time processing capabilities, which can be challenging to achieve without compromising accuracy

VII. Metrics, Objectives, and Business Goals

Metrics

- ➢ Accuracy: This metric measures the overall correctness of the model across all classes. It is calculated as the ratio of correctly predicted observations to the total observations. Accuracy is useful when the classes are well balanced but can be misleading if there's a significant class imbalance.
- ➢ Recall: Also known as sensitivity, this metric calculates how well the model can identify positive cases among all actual positives. It is defined as the ratio of true positives to the sum of true positives and false negatives. Recall is crucial when the cost of missing a positive prediction is high.

Objectives

- ➢ Develop a highly accurate model for brain tumor detection and classification.
- ➢ Reduce diagnostic time and support radiologists with reliable AI tools.
- ➢ Enhance patient outcomes through early and precise tumor identification.

VIII. Failure Analysis

Potential risks include model overfitting, data privacy breaches, and hardware failures during deployment. Mitigation strategies involve extensive validation, regular audits of data handling practices, and robust backup systems for deployment infrastructure.

- **Misclassification of Non-Tumor Conditions**: Certain diseases or abnormalities may present in MRI scans with appearances similar to brain tumors. This similarity can lead to false positives, where the model incorrectly classifies these conditions as brain tumors.
- **Corrupted File Uploads**: If a corrupted image file is uploaded, the model might fail to process the image correctly or might produce unreliable outputs. Ensuring that the model can handle or reject corrupted files is essential.
- **Inappropriate MRI Uploads**: Uploading an MRI that is not relevant (e.g., an MRI of a different body part or a non-diagnostic quality image) can lead to incorrect predictions. Implementing a pre-processing step to validate that the uploaded MRI is appropriate for analysis might mitigate this issue.
- **Poor Image Quality**: Extremely low-quality images can significantly impair the model’s ability to accurately analyze and classify the content. This could be due to factors such as low resolution, high noise levels, or poor contrast. Establishing quality control measures to either enhance the image quality or reject inadequate images can be beneficial.

IX. Deployment Infrastructure

Project Development on Google Cloud Platform (GCP)

The entire project will be developed on Google Cloud Platform (GCP), chosen for its substantial benefits which include:

- **Scalability**: GCP provides dynamic scaling options that can efficiently handle the varying load of MRI data processing and analysis, accommodating growth without compromising performance.
- **Reliability**: With its robust infrastructure and uptime guarantees, GCP ensures that the application remains operational and accessible, minimizing downtime.
- **Ease of Use**: GCP offers an intuitive interface and comprehensive documentation that simplifies setup, deployment, and management of cloud resources.
- **Tool Support**: It supports a wide range of tools and services that can be integrated seamlessly, catering to diverse needs from data storage to machine learning and analytics.

Frontend Development

The frontend of the project will be developed using Streamlit, which is compatible with Linux, macOS, and Windows. Streamlit allows for quick creation of interactive dashboards, making it suitable for visualizing MRI data analysis and model outputs in a user-friendly manner.

Backend Infrastructure

The backend, where the model will be served and monitored, will also be hosted on GCP. This setup ensures that the model is scalable, secure, and integrated with GCP's powerful computing resources.

Key GCP Tools Utilized

Several specific GCP tools will be employed to optimize performance and functionality:

- **Buckets**: Used for scalable and secure storage of MRI images and other data.
- **Vertex AI**: Provides a managed machine learning service that aids in training, tuning, and deploying the model directly on GCP.
- **Data Studio**: Utilized for creating interactive dashboards. This tool will help in visualizing the model's performance metrics and operational analytics, making it easier to monitor and refine the processes.

X. Monitoring Plan

1. Statistics and Metadata of the Image Received for Prediction:

Purpose: To ensure the quality and appropriateness of each MRI image uploaded for analysis. This involves collecting and analyzing metadata such as image resolution, file size, and format.

Benefits: Monitoring these aspects helps in detecting any anomalies or deviations from expected standards, such as low-resolution images or unsupported formats, which could affect the model's accuracy. Ensuring high-quality images allows the model to make more accurate predictions, reducing the likelihood of false negatives or positives.

2. Metrics Representing Health of the Model:

Purpose: To continuously assess the model's performance and operational status. Key performance indicators might include accuracy, precision, recall, and F1-score. Additionally, monitoring other metrics like confusion matrix, ROC-AUC score, and log loss can provide a more comprehensive view of model health.

Benefits: Regularly evaluating these metrics ensures that the model maintains its efficacy over time and remains reliable. Any sudden changes in these metrics can indicate potential issues needing investigation or intervention. For example, a drop in precision might indicate an increase in false positives, prompting a review of the data or model parameters.

3. Monitoring for Dataset Shift and Dataset Skew:

Dataset Shift: Occurs when the statistical properties of incoming data differ from the training data. This can affect the model's performance if not addressed. For instance, changes in imaging techniques or patient demographics could lead to a dataset shift.

Dataset Skew: Refers to when the model receives data that is not representative of the population data it was trained on, potentially leading to biased predictions. This might occur due to seasonal variations or changes in the patient population over time.

Purpose: To detect and correct for shifts or skews in the data which could degrade the model's accuracy and fairness. This includes implementing techniques to regularly compare incoming data distributions with training data distributions and identify significant deviations.

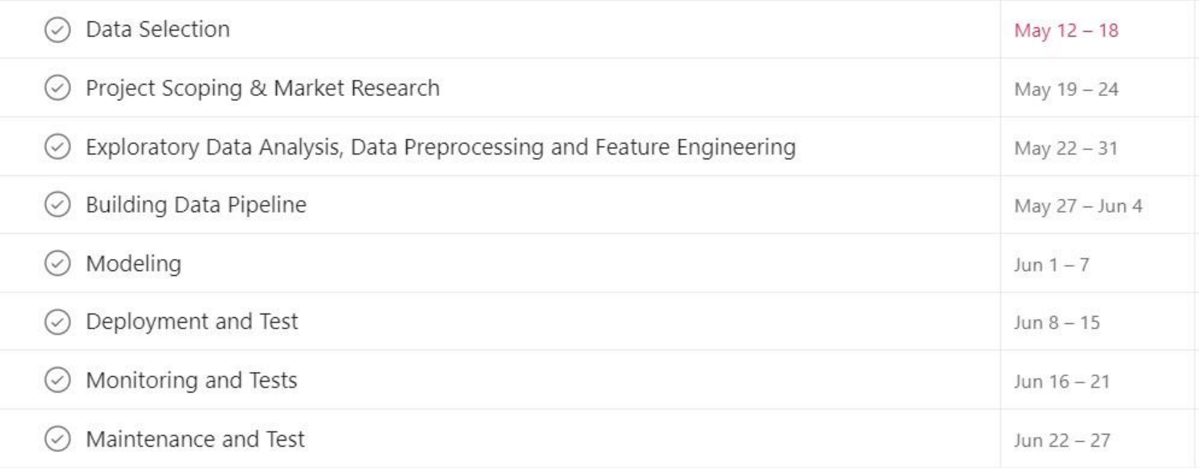
Benefits Proactive monitoring for these issues allows for timely adjustments to the model or its training data, thus maintaining the model's accuracy and generalization capabilities. Early detection of dataset shifts or skews can prompt actions such as model retraining, data augmentation, or updating preprocessing steps to ensure the model adapts to new patterns in the data.

X. Success and Acceptance Criteria

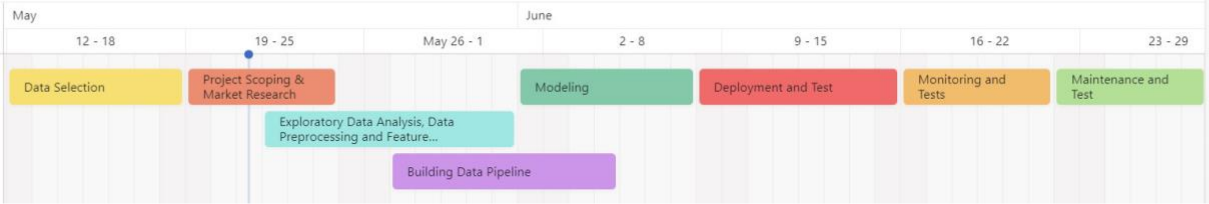
1. **Good Feedback from Doctors About the Performance of the Model**:
   - **Definition**: Collect qualitative feedback from medical professionals who interact with the model. This feedback should focus on the accuracy, reliability, and usability of the model in clinical settings.
   - **Measurement**: Implement surveys or interviews to gather detailed feedback. Key points could include satisfaction with the model's diagnostic support, the clarity of the model's predictions, and its integration into clinical workflows.
   - **Success Threshold**: A high level of satisfaction (e.g., above 80% positive feedback) would be indicative of success.

2. **End User Trust**:
   - **Definition**: End user trust involves users' confidence in the model'spredictions and their willingness to rely on it for making clinical decisions.
   - **Measurement**: Assess trust through direct feedback and indirect indicators such as the rate of follow-through on the model's recommendations. Another method is to analyze the frequency of override or disregard for the model's advice.
   - **Success Threshold**: Establish benchmarks for levels of trust and acceptance, such as less than 20% override rates and high reliance on the system for decision-making in relevant cases.

XI. Timeline Planning



Data Selection	May 12 – 18
Project Scoping & Market Research	May 19 – 24
Exploratory Data Analysis, Data Preprocessing and Feature Engineering	May 22 – 31
Building Data Pipeline	May 27 – Jun 4
Modeling	Jun 1 – 7
Deployment and Test	Jun 8 – 15
Monitoring and Tests	Jun 16 – 21
Maintenance and Test	Jun 22 – 27



I. Additional Information

The methodologies for data preprocessing, modeling, deployment, and monitoring described above could change throughout the development process. Any changes will be documented and the corresponding documentation will be updated to reflect these modifications