Deep Learning Project: Predicting Recovery Time from MRI Data

1. Project Objective

The primary aim of this project is to develop a predictive model that estimates the recovery time (in days) for patients based on their blood flow MRI (data and demographic/medical features. By achieving this, we aim to provide valuable insights into patient recovery and improve medical decision-making processes.

2. Problem Statement

The key question driving this project is: **Can we accurately predict recovery time based on blood flow MRI data, demographic features (e.g., age, gender), and medical history (e.g., history of concussion)?**

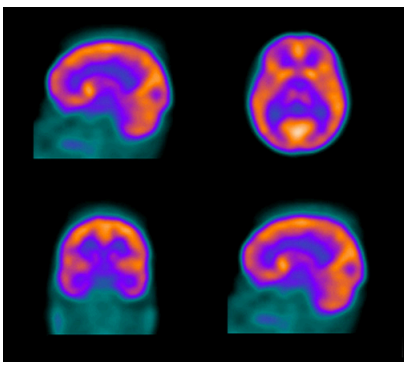
This project involves leveraging a deep learning approach to analyze blood flow MRI images and other structured data, enabling the development of a robust predictive tool.

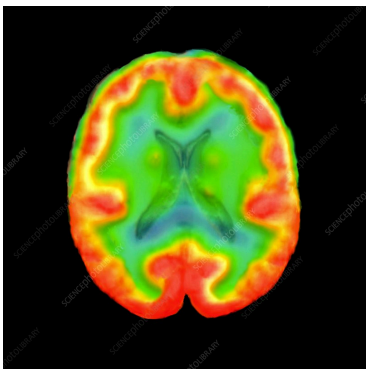
3. Data Overview

The dataset comprises MRI image data and tabular features for patients. The details are as follows:

- **Blood flow MRI Data**: High-dimensional data representing brain scans. Images of blood flow to certain areas of the brain.
- **Demographic Features**:
    - **Age**: Continuous variable.
    - **Gender**: Binary variable (0: Female, 1: Male).
- **Medical History**:
    - **History of Concussion (HOC)**: Binary variable (0: No history, 1: History).
- **Days After Injury**: The number of days after the injury when the MRI was taken.
- **Target Variable**: **Days of Recovery** (continuous variable).

Below are visualizations of MRI data:





Below is a summary of the dataset:

Feature	Type	Description
Age	Continuous	Age of the patient
Gender	Binary	Gender of the patient (0: Female, 1: Male)
HOC	Binary	History of concussion
Days After Injury	Continuous	Time elapsed since the injury (in days)
fMRI Data	Image	High-dimensional image data
Days of Recovery (y)	Continuous	Target variable to predict

4. Tools and Techniques

Converting fMRI Data into a Numerical Array

- **HDF5 and h5py**:
  - **Purpose**: The HDF5 format is a hierarchical data format commonly used for storing large-scale scientific data, like fMRI scans.
- **SimpleITK**:
  - SimpleITK is another Python library often used for working with medical images, particularly in radiology and fMRI.

Deep Learning Framework: PyTorch

PyTorch is chosen for its flexibility, ease of debugging, and dynamic computation graph. These characteristics make it particularly suitable for iterative model development.

Neural Network Model

- **Cross-Validation**

  5-fold cross-validation to evaluate model generalizability.

Convolutional Neural Network (CNN)

- **Spatial Relationships**:
  - fMRI data (or MRI images) contain spatial patterns and structural information about the brain. CNNs are excellent at capturing these spatial hierarchies through convolutional layers.
- **Dimensionality Reduction**:
  - Instead of flattening the fMRI images into large arrays (as with fully connected layers), CNNs preserve the spatial structure and reduce dimensionality efficiently.
- **Feature Extraction**:

- CNNs automatically learn spatial features such as edges, textures, and patterns in the brain data that may be predictive of recovery time.

Why is this method appropriate for this dataset?

- **Understanding the Data Structure**:
  - The dataset includes high-dimensional MRI data, which contains spatial features that are essential for predicting recovery time.
  - CNNs are well-suited for handling **spatial dependencies** in images, making them an ideal choice for MRI data rather than fully connected networks that lose spatial relationships.
- **Why CNNs Work Well for MRI Data?**
  - CNNs leverage **convolutions** to capture important brain features that influence recovery time.
  - They apply **filters** to detect structural differences in brain regions, which are crucial for medical predictions.
- **Why Cross-Validation?**
  - MRI data can be **noisy** and **limited in size**. Using **5-fold cross-validation** ensures that the model generalizes well and does not overfit to a specific subset of the data.

5. Project Timeline

The following Gantt chart outlines the major milestones of the project: