**Project Proposal DS 3002 Data Mining**

**Brain Tumor MRI Classification**

**Group Members:**

Muhammad Ahmad(21L-5617)

Nimra Amer(21L-5609)

* **Problem and DataSet Description:**

In this project, we will focus on the problem of “Multiclass Classification” of the type of brain tumor in images. For this purpose, the image dataset we are using is collected from Kaggle <https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset/data>

The dataset consists of approx. 7022 images. The dataset consists of four classes of brain tumors that are “Glimo”, “Meningioma”, “Notumor” and “Pituitary”. Each image belongs to one of these classes.

The Primary Objective of this project is to develop and evaluate machine learning **and deep learning** models for the automated multi-class detection and classification of brain tumors from MRI images. Specifically, we aim to explore the effectiveness of logistic regression, basic Convolutional Neural Networks (CNNs), and **pre-trained** CNN models such as vgg-19**.**

* **Models/Algorithms/Techniques:**

We will specifically be using Logistic Regression from machine learning models, CNN from deep learning models, and pre-trained CNN models such as vgg-19.

**Pre-Processing Techniques:**

Before implementing any of the above models, the first thing we will do is the pre-processing of images. The pre-processing techniques may include(can be fewer/more in actual implementation):

* Image Resizing
* Intensity Normalization
* Noise Reduction
* Histogram Equalization
* Skull Stripping
* Image Registration
* Data Augmentation
* Region of Interest (ROI) Cropping
* Dimensionality Reduction
* Quality Control

**Implementation Approach for Models:**

1. We will begin with logistic regression as a baseline model. This model will help us differentiate between different types of brain tumors like gliomas, meningiomas, pituitary tumors, and acoustic neuromas. By doing so, we can establish a basic understanding of the classification task and set a benchmark to evaluate more complex models.
2. Next, we will create and train a basic CNN architecture specifically designed for detecting multiple types of brain tumors. This CNN will be designed to capture complex spatial patterns and subtle features within MRI images, thereby enabling precise classification across various tumor categories. To improve the model's strength and generalization capabilities, we will utilize techniques such as data augmentation and regularization.
3. After that, We will be using transfer learning techniques in our project to improve brain tumor detection across different tumor categories. This involves using pre-trained CNN models like VGG-19 and fine-tuning them on our medical imaging dataset. By doing so, we can leverage pre-existing knowledge from large-scale image datasets such as ImageNet to enhance the classification performance.

The main purpose of using multiple models will be to increase and improve the accuracy of our models.

* **Software Tools:**

We will be coding in Python. The libraries that we will be using may include(not limited to) sklearn, Tensorflow/pytorch, Keras, OpenCV, Seaborn/ Matplotlib, etc.

* **Expected Results And Evaluation:**

After implementing all the models, we expect to get the results for multiclass classification by all the models. Furthermore, we will be evaluating the models based on their accuracy, f1-score, precision, recall, etc.

* **Preliminary Result and Dataset Explored:**

In this brain tumor diagnosis project, a dataset comprising 7023 MRI images of human brains is utilized, combining three separate datasets. These images are categorized into four classes: glioma, meningioma, no tumor, and pituitary. The "no tumor" class images were sourced from the Br35H dataset, while issues were identified with the glioma class images in the SARTAJ dataset, prompting their exclusion from the analysis. Instead, images from the Figshare site were incorporated to rectify this issue.

Following prior research and personal experimentation with various models, it was determined that the original SARTAJ dataset exhibited inconsistencies. Consequently, the decision was made to substitute the glioma class images to ensure the accuracy and reliability of subsequent analyses and model training processes (Masoud NickParvar, 2021).

* **Outline of the work-to-do:**
* Loading all the images, decoding them into pixel arrays, to create a vector of images from the dataset.
* Applying different pre-processing techniques on the images.
* Applying the first model to get the predictions and its accuracy.
* Repeat the third step for all the remaining models.
* Comparing the models’ performance based on classification report.