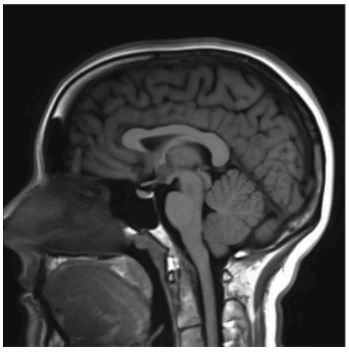
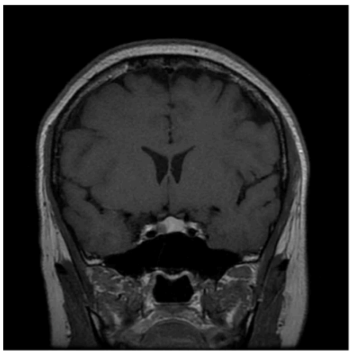
Processus Data - Final Project

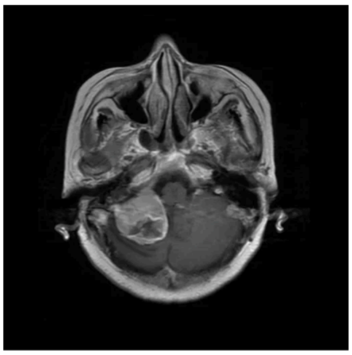
Objective: Develop an efficient and scalable MLOps pipeline for automating the machine learning workflow



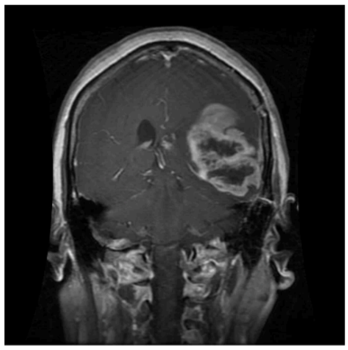
notumor



pituitary



meningioma



glioma

Réalisé et présenté par :

Zakaria KOUHIL
Yasmine ZEMMOURI
Ilyas GHAZOUI
Yasmine MESDOUR

Encadré par :

Feda ALMUHISEN

amU
Aix Marseille Université

1. Introduction

The main goal of this project is to create a reliable MLOps pipeline for classifying brain tumours using MRI scans. The "Brain Tumor MRI Dataset" from Kaggle was used, containing MRI images labelled as either normal or tumour, with tumour subcategories such as glioma, meningioma, and pituitary tumours. This project demonstrates the use of MLOps tools for tracking experiments, improving models, and managing deployment and monitoring.

This report explains the methods used, challenges faced, and results achieved during the development and deployment of machine learning models for brain tumour classification. The focus is on making the process efficient, reproducible, and scalable.

2. Data Exploration

The dataset includes 7,023 MRI images divided into four categories: Glioma Tumour, Meningioma Tumour, Pituitary Tumour, and No Tumour. Each image was resized to 128x128 pixels and normalised to ensure consistent inputs for training. The dataset was initially imbalanced, with 1,621 Glioma images, 1,645 Meningioma images, 1,757 Pituitary images, and 2,000 No Tumour images. To address this imbalance, data augmentation techniques such as image rotation were applied to increase the number of samples in the underrepresented classes, creating a more balanced dataset.

The data was preprocessed by scaling pixel values to the range $[0, 1]$ and converting target labels into one-hot encoded format for multi-class classification. Tumour images displayed distinct patterns, making them well-suited for feature extraction by deep learning models. The "No Tumour" category presented challenges due to noise and artefacts, requiring robust preprocessing to ensure accuracy. Overall, the augmented and preprocessed dataset provided a strong foundation for building reliable and interpretable machine learning models.

3. Model Development

Four models were created for the task of classifying brain tumours. **MobileNetV2**, a pre-trained transfer learning model, was fine-tuned with additional layers for this specific dataset. It effectively used its pre-trained features to identify patterns in the images. A **Shallow Neural Network** with three hidden layers was developed as a simpler, lightweight model for comparison.

A **Custom Convolutional Neural Network (CNN)** was built with three convolutional layers, followed by max-pooling and fully connected layers. It allowed adjustments to hyper-parameters such as the dropout rate and optimisers. A **Random Forest Classifier** was also used as a traditional machine learning approach, providing high interpretability and computational efficiency.

The **Random Forest** model performed best, achieving 92% accuracy, alongside MobileNetV2, which also scored 92%. However, Random Forest was chosen as the best model due to its simplicity, interpretability, and strong performance. The Custom CNN scored 95% accuracy, while the Shallow Neural Network achieved 84%. Although the CNN had the highest accuracy, its complexity made Random Forest more practical for deployment.

4. MLOps Approach

MLOps practices were applied to manage the entire lifecycle of the machine learning models. Tools such as **MLflow**, **LIME**, and **Flask** were used to track experiments, interpret model outputs, and deploy predictions.

- **Experiment Tracking**: MLflow was used to log hyper-parameters, metrics like accuracy and recall, and model files. This made it easy to compare and reproduce experiments.
- **Model Interpretability**: LIME (Local Interpretable Model-agnostic Explanations) was used to explain individual predictions, helping users understand the decisions made by the models.
- **API Integration**: A Flask API was developed to deploy the Random Forest model. Users can upload MRI images through the API and receive real-time predictions with explanation insights from LIME.
- **Model Versioning and Deployment**: MLflow enabled versioning and storage of model files, ensuring that the best model was reliably deployed. Flask acted as the interface for serving predictions, making the solution accessible to end-users.

By combining these tools, the project achieved a robust and user-friendly machine learning pipeline for real-world applications.

5. Conclusion

This project successfully created a reliable pipeline for classifying brain tumours using MRI scans. The Random Forest model was identified as the best-performing due to its accuracy, simplicity, and ease of use. MLOps tools like MLflow ensured that all experiments were tracked and reproducible, while LIME provided transparency into the model's predictions.

The Random Forest model was deployed through a Flask API, enabling users to upload images and receive predictions quickly. This project shows how MLOps practices can streamline machine learning workflows and deliver scalable, interpretable, and practical solutions. Future work could include expanding the dataset, automating retraining, and deploying the API at a larger scale.