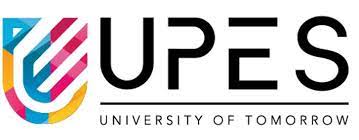
****

**Assignment**

**Course & Batch: BTech CSE AIML (B-5)**

**Student Details:**

**Name: Lakshya Agrawal**

**Sap ID: 500105927**

**Roll No. – R2142220338**

**Brain Scan Dataset Report**

**1. Introduction**

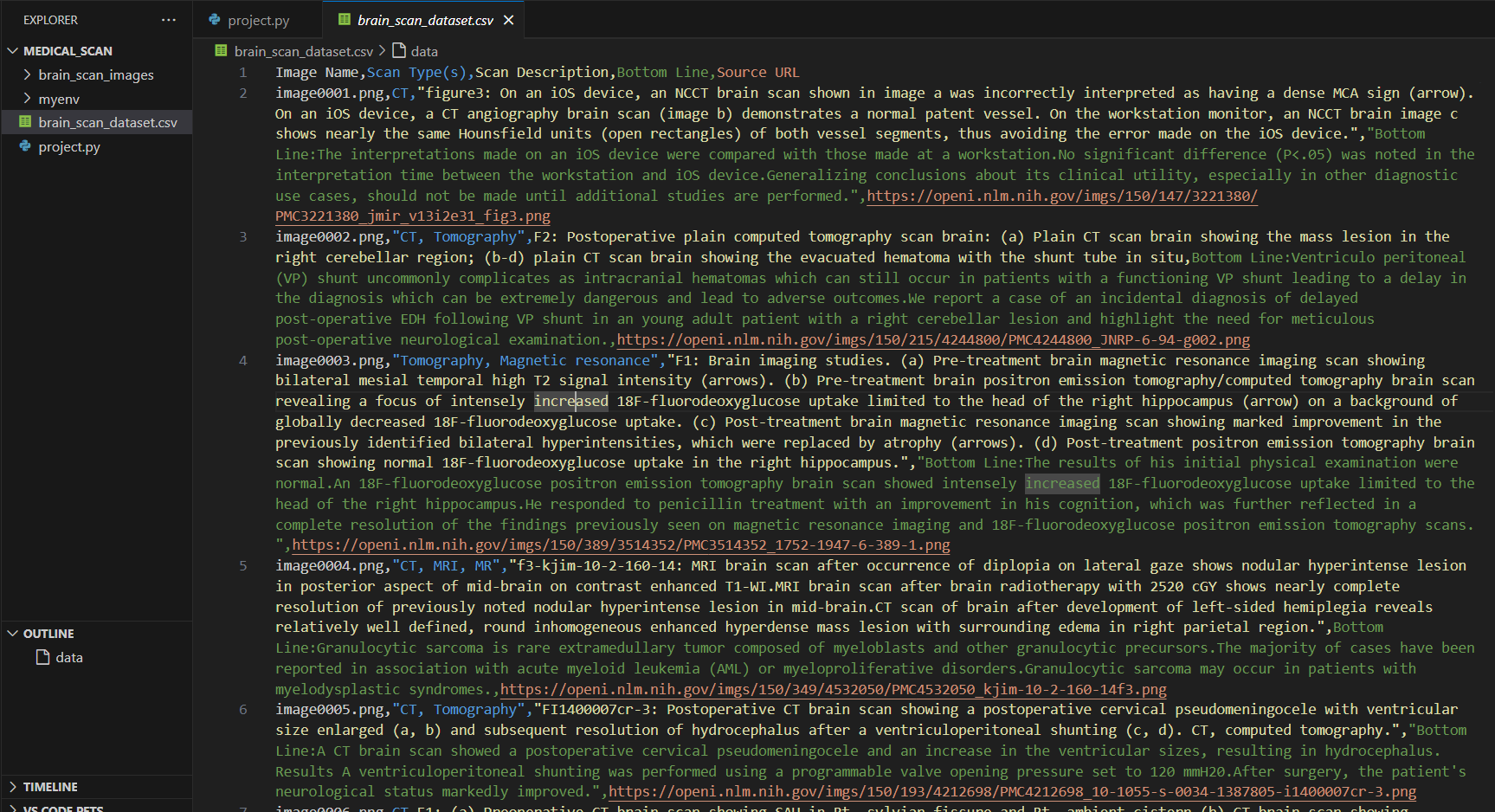
The **Brain Scan Dataset** is a curated collection of brain imaging data sourced from **OpenI (National Library of Medicine)**, containing various types of brain scans such as **CT, MRI, X-ray, PET, Ultrasound, and more**. This dataset was developed for **machine learning applications in medical image analysis**, aiding in **disease detection, segmentation, and classification.**

**2. Data Collection Process**

* **Source:** OpenI medical image repository
* **Collection Method:** Web scraping using Selenium and BeautifulSoup
* **Number of Images Collected:** 1,000
* **Total Pages Scraped:** 12 pages (100 images per page)
* **Metadata Extracted:**
  + Image Name
  + Scan Type(s)
  + Scan Description
  + Bottom Line Summary
  + Source URL

A screen shot of a computer program

AI-generated content may be incorrect.



**3. Dataset Statistics**

**Total Images: 1000**

**Scan Types Distribution:**

**CT: 681**

**MRI: 227**

**X-ray: 5**

**PET: 51**

**Ultrasound: 5**

**Microscopy: 5**

**Graphics: 0**

**Description Availability:**

**Fully annotated images: 1000**

**Partially annotated images: 0**

**Bottom Line Summary Availability:**

**Present in: 818**

**Absent in: 182 images**

A screenshot of a computer screen

AI-generated content may be incorrect.

**4. Data Preprocessing**

The images were processed before use in machine learning models:

* **Resolution normalization:** All images were resized to **150x150 pixels**.
* **Format standardization:** Saved as **PNG** files.
* **Noise reduction:** Applied Gaussian filtering to improve clarity.
* **Metadata structuring:** Extracted details organized in CSV format.

**5. Machine Learning Applications**

This dataset can be used for various **medical AI applications**, including:

* **Image Classification:** Predicting scan type (CT, MRI, X-ray, etc.) using CNNs.
* **Disease Detection:** Identifying abnormalities using deep learning models like **ResNet, EfficientNet, or UNet.**
* **Segmentation Tasks:** Training UNet for tumor detection and segmentation.
* **Multi-Label Classification:** Predicting multiple conditions from scan descriptions using **BERT-based NLP models.**

**6. Challenges & Limitations**

* **Incomplete Annotations:** Some images lack descriptions or scan types.
* **Class Imbalance:** CT & MRI scans are overrepresented.
* **Diverse Image Quality:** Variations in resolution and contrast among images.
* **Lack of Ground Truth Labels:** No direct confirmation of disease presence.

**7. Conclusion**

The **Brain Scan Dataset** serves as a valuable resource for AI-driven **medical imaging research**. Future improvements could include **manual verification of annotations, expansion to other scan types, and inclusion of structured disease labels.** This dataset will aid in **automating radiological diagnosis, reducing workload for radiologists, and improving patient outcomes.**

**8. References**

1. OpenI - National Library of Medicine ([https://openi.nlm.nih.gov](https://openi.nlm.nih.gov/))
2. Ronneberger, Fischer, Brox - **"U-Net: Convolutional Networks for Biomedical Image Segmentation"**
3. He, Zhang, Ren, Sun - **"Deep Residual Learning for Image Recognition" (ResNet)**
4. Vaswani et al. - **"Attention is All You Need" (Transformer-based NLP for Medical Data)**