Home Assignment: Data Scientist | Samueli Institute

Part 1: Deep Learning for Clinical Neuro-Oncology

Clinical Context

In neuro-oncology, the early detection of **intraparenchymal brain metastases** is a cornerstone of clinical decision-making, directly influencing treatment pathways such as neurosurgery, targeted radiation, or systemic therapies. Although MRI remains the gold standard for definitive diagnosis, **Non-Contrast CT (NCCT)** is frequently the frontline modality in emergency and screening settings due to its rapid acquisition and widespread availability.

The Challenge

Identifying intraparenchymal metastases on NCCT scans is notoriously difficult. Unlike Contrast-Enhanced CT, where lesions typically demonstrate prominent enhancement ("lighting up"), findings on NCCT can be extremely subtle. Often, the only indicators are slight variations in tissue density (**hypodensity**) or secondary clinical signs such as **peritumoral edema** (localized swelling) or **mass effect** (displacement of anatomical brain structures)

Your Goal: Develop a robust deep learning model to accurately flag CT slices containing these metastatic lesions.

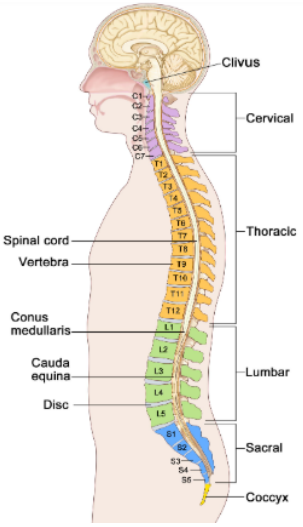
Dataset

The data for this task is available at the following link: brain_metastasis

- **Images**: Approximately 5,000 2D NCCT slices in **DICOM** format.
- **Labels**: A corresponding CSV file providing the ground truth for each slice:
    - **Label 1**: Presence of an intraparenchymal metastasis.
    - **Label 0**: Normal scan (no detectable pathology).

Part 2: Automated Localization of the L3 Vertebra

Clinical Significance

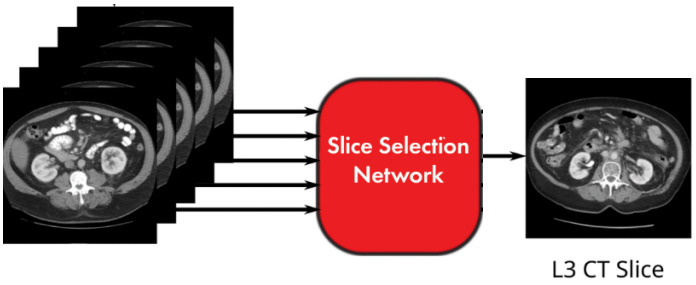


The lumbar spine (L1–L5, in green) is a fundamental structural component of the human skeleton. In the field of **Body Composition Analysis**, the mid-axial slice of the **L3 vertebra** has been established as a standardized anatomical landmark.

Measuring tissue distribution at the L3 level—specifically skeletal muscle mass and adipose tissue—provides a highly accurate proxy for whole-body composition. This analysis is a critical tool for identifying conditions such as **sarcopenia** (muscle wasting) and cachexia, which are significant prognostic factors in oncological and geriatric care.

The Challenge

Given a full 3D CT volume (consisting of a series of DICOM files), your task is to design a model that identifies and returns the specific slice index corresponding to the **middle of the L3 vertebra**.



Please propose **two different architectural approaches** to solve this problem. For each suggestion, provide a detailed description covering the following aspects:

1. **Input & Output Definitions:** What exactly enters the model and what is the mathematical form of the prediction?
2. **Preprocessing Pipeline:** Describe the necessary steps to prepare the raw DICOM data.
3. **Model Architecture:** Detail the type of network you would use
4. **Workflow Steps:** A step-by-step breakdown of the process from raw input to the final slice selection.
5. **Visuals/Pseudo-code:** You are encouraged to use diagrams, flowcharts, or pseudo-code to illustrate your logic (actual implementation code is not required).

Submission Requirements

To complete the assignment, please provide a comprehensive submission covering both Part 1 and Part 2. Your submission should include the following:

1. Technical Implementation (GitHub Repository)

A link to a private or public GitHub repository containing:

- **Source Code:** Fully developed, clean, and modular code for part 1 (Neuro-Oncology).
- **Documentation:** A `README.md` file with instructions on how to set up the environment and execute the training/inference pipelines.
- **Reproducibility:** Ensure all dependencies are listed (e.g., `requirements.txt` or `environment.yml`).

2. Executive Summary & Technical Presentation (PDF/PowerPoint)

A structured presentation summarizing your work. This document should be designed as if you are presenting your findings to the Samueli Institute's lead researchers. Please include:

- **Methodology & Rationale:** A clear articulation of your chosen architectures and why they were selected for these specific clinical tasks.
- **Experimental Design:** Details on preprocessing (windowing, resampling), data augmentation, and training strategies.
- **Performance Evaluation:** A deep dive into the outcomes using relevant metrics (e.g., Sensitivity, Specificity, AUC-ROC, or localization error).
- **Challenges & Insights:** A discussion on the obstacles encountered (e.g., class imbalance, DICOM complexities) and how you addressed them.
- The focus is on **your reasoning and experimental design**, not on extensive fine-tuning or achieving the highest possible performance.

For any question: hillavar@clalit.org.il

Good Luck!