**Title: Develop a Chat-GPT like Large Vision-Language Model for Breast Cancer Malignancy Prediction**

The primary objective of this project is to implement a ChatGPT like [Large Vision-Language Model (VLM)](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10445007) for application to medical diagnosis. Particularly, the focus is on Breast Cancer which is one of the leading causes of cancer-relate mortality among women worldwide. This makes early making early detection of [Breast Cancer Malignancy Prediction](0133840786_Ch11-21-231272.ziphttps:/journals.lww.com/md-journal/fulltext/2024/01190/breast_cancer__a_review_of_risk_factors_and.67.aspx) of significant importance. In the envision LVM, two distinct representations are derived, one from the [Deep Learning (DL)](https://www.nature.com/articles/nature14539)-based Text Encoder, and one from a DL-based Image Encoder using Transformer Architecture (which is the main module of a VLM such as ChatGPT). The image encoder will be trained based on multi-view X-ray images (mammograms). The extracted representations (also referred toas embeddings) will serve as the key components for the task of malignancy prediction.

**Project Scope**

The proposed model will use one of the most revolutionary LVM models, i.e., the [Contrastive Language-Image Pre-training (CLIP) model](https://arxiv.org/pdf/2103.00020). The variant of CLIP to be implemented through this project will encode mammogram images and textual information based on radiologists’ diagnosis into a shared latent space to analyze mammograms based on te similarity to their associated clinical descriptions.

**Inputs:**

The [CBIS-DDSM](https://www.cancerimagingarchive.net/collection/cbis-ddsm/) **(Curated Breast Imaging Subset of DDSM)** is a publicly available dataset designed for breast cancer **research. It is an updated and standardized version** of the **Digital Database for Screening Mammography (DDSM).**

**(Phase I) Tasks:**

* Loading the [pre-trained CLIP model](https://arxiv.org/abs/2103.00020), including both its **text encoder** and **image encoder**
* Constructing text prompts for training, i.e., format the class labels into a structured text prompt such as “A photo of a [class label],” using ChatGPT.
* Use these generated sentences as input to the text encoder during training.
* Keep the image encoder and text encoder frozen to preserve their pre-trained feature extraction capabilities.
* Modify the model architecture to integrate additional Fully Connected (FC) layers (fusion layers) after the image encoder’s output for learning domain-specific representations.
* Train these layers to adjust their weights based on the similarity between text and image features.
* Once trained, remove the text encoder from the inference pipeline.
* At test time, only the image encoder and the trained FC layers are used for classification.

**Expected Output:**

The model should classify each input image into one of two categories Benign/Malignant Mass through quantitative metrics (e.g., accuracy, recall,…)