**Dataset Summary**

* **Dataset Name:** Breast-Cancer-Screening-DBT | Breast Cancer Screening - Digital Breast Tomosynthesis.
* **Dataset Link:** [BREAST-CANCER-SCREENING-DBT - The Cancer Imaging Archive (TCIA)](https://www.cancerimagingarchive.net/collection/breast-cancer-screening-dbt/).
* **Data Source and Collection:**

The dataset was obtained from the Duke Health System, with data collected between January 1, 2014, and January 30, 2018. DBT volumes were acquired, and radiology reports containing the term "tomosynthesis" and pathology reports containing the term "breast" were used to identify relevant cases.

* **Data Groups:**

The dataset is divided into four groups:

* 1. Normal: Includes screening studies without abnormal findings.
  2. Actionable: Includes studies that led to further imaging due to noted abnormalities.
  3. Benign: Contains studies with biopsy-proven benign masses or architectural distortions.
  4. Cancer: Contains studies with biopsy-proven cancerous masses or architectural distortions.
* **Data Split:**

The dataset consists of a total of 22,032 reconstructed volumes from 5,610 studies and 5,060 patients. It was randomly split into training, validation, and test sets, ensuring no overlap of patients between subsets.

* **Image Annotation:**

Radiologists annotated the images by identifying and enclosing masses and architectural distortions in the central slice of each DBT volume with rectangular bounding boxes. These annotations were performed using custom software developed by the researchers.

* **Radiologist Expertise:**

Annotations were conducted by two experienced radiologists, with 25 and 18 years of experience in breast imaging, respectively.

* **Final Notes:**

Due to the limited resources (GPUs) and time, training our model using the whole dataset (a total size of **1.63TB**) would be impossible, that is why we would use the training dataset only (a total size of **135.77GB**) in our training & testing phases, however this may affect the performance of our model due to the limited amount of data available for training.

* **More information about the dataset can be found in the referenced paper below:**

M. Buda *et al.*, “A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images,” *JAMA Network Open*, vol. 4, no. 8, p. e2119100, Aug. 2021, doi: 10.1001/jamanetworkopen.2021.19100.