Phase-2: Preprocessing & Dataset Preparation

Objective:
Prepare a fully preprocessed, balanced mammography dataset that can be used for any downstream model (image-only, metadata-only, or multimodal), without committing to any train/test split prematurely.

---

1. Dataset Categorization

- **Case Categories Defined:**
    - **0:** Both breasts normal
    - **1:** Cancer in one breast
    - **2:** Cancer in both breasts
- **Implementation:**
    - Derived from `birads_left` and `birads_right` columns.
    - Added a new column `case_category` in `image_df` for image-level labeling.

Counts Before Upsampling (study-level):

0 (Both breasts normal): 3614 studies
1 (Cancer in one breast): 295 studies
2 (Cancer in both breasts): 90 studies

-

---

2. Upsampling Strategy

- **Goal:** Address class imbalance at **study/patient level**.

- **Approach:**
    - Duplicate entire studies for minority classes (cases 1 and 2) while keeping all associated images intact.
    - Only applied on the **training split**, not test split.

Study-level Counts After Upsampling:

0 (Normal): 3614 studies
1 (One breast cancer): 2295 studies
2 (Both breasts cancer): 1090 studies

-

Image-level Counts After Upsampling:

0 (Normal): 18072 images
1 (One breast cancer): 15928 images
2 (Both breasts cancer): 14908 images

-

---

3. Image Preprocessing Pipeline

All images (original + upsampled) underwent:

1. **Breast Region Detection (BRD):**
    - Gaussian blur (5×5 kernel)
    - OTSU thresholding
    - Contour detection → select largest contour → crop breast region
2. **Contrast Enhancement:**
    - CLAHE (Contrast Limited Adaptive Histogram Equalization)
3. **Truncated Normalization:**

    - Clip to 5th–99th percentiles
    - Normalize to 0–255
- **Implemented as a single function:** `preprocess_mammogram(image_path)`
- **Verified visually and quantitatively:**
    - Sample previews of normal/abnormal images
    - Image-level statistics: width, height, mode (`L` → grayscale)

---

4. Folder Structure for Preprocessed Dataset

```
birads_preprocessed_dataset/
│
├── training/
│   ├── normal/
│   │   └── Study_<study_id>/
│   │       └── Image_<image_id>.png
│   └── abnormal/
│       └── Study_<study_id>/
│           └── Image_<image_id>.png
│
└── test/
    ├── normal/
    └── abnormal/
```

- Ensures **study-level grouping** is preserved.
- Preprocessed images retain their **original grayscale format and dimensions**.
- Ready for downstream model training/evaluation.

---

5. Dataset Summary

Split	Label	Studies	Images

training	normal	6782	20102
training	abnormal	3730	7894
test	normal	995	3616
test	abnormal	187	384

- **Total images:** 31,996
- **Width:** 436–912 px, **Height:** 1019–1520 px

---

6. Dataset Storage

- Preprocessed dataset saved to local Colab folder:
  `/content/birads_preprocessed_dataset`

Can be **zipped with folder structure intact** and copied to Google Drive for download:

/content/drive/MyDrive/birads_preprocessed_dataset.zip

Phase-2 Outcome:
A fully preprocessed, study-level balanced mammography dataset with train/test splits maintained, ready for any downstream modeling, with robust image preprocessing and upsampling applied only to the training set.