Developing a ChatGPT-like Vision-Language Model for Breast Cancer Malignancy Prediction

Capstone Project | DSCI 592

Team Members: Aditya Sinha, Ram Kishore, Sanskruti Chavanke, Samriddhi Singh, Sean Smyth, David Lin

Introduction

- ❖ Build a Vision-Language Model (VLM) based on the CLIP architecture tailored for breast cancer malignancy prediction.
- ❖ Integrate multimodal data (mammogram images + clinical text) into a shared latent space.
- ❖ Improve diagnostic accuracy for early detection of malignant masses.
- ❖ Develop a lightweight, deployable model for potential real-time clinical use.

Dataset Overview

- **Dataset:** CBIS-DDSM (Curated Breast Imaging Subset of the DDSM).
- **Images:** Over 3,000 mammograms with labels (Benign / Malignant).
- **Metadata:** Includes image resolution, body part examined, modality type, and clinical notes.
- **Image Types:** Cropped images, full mammograms, and ROI masks.
- **Goal:** Preprocess and balance dataset for robust model training.

Dataset Cleaning & Preprocessing

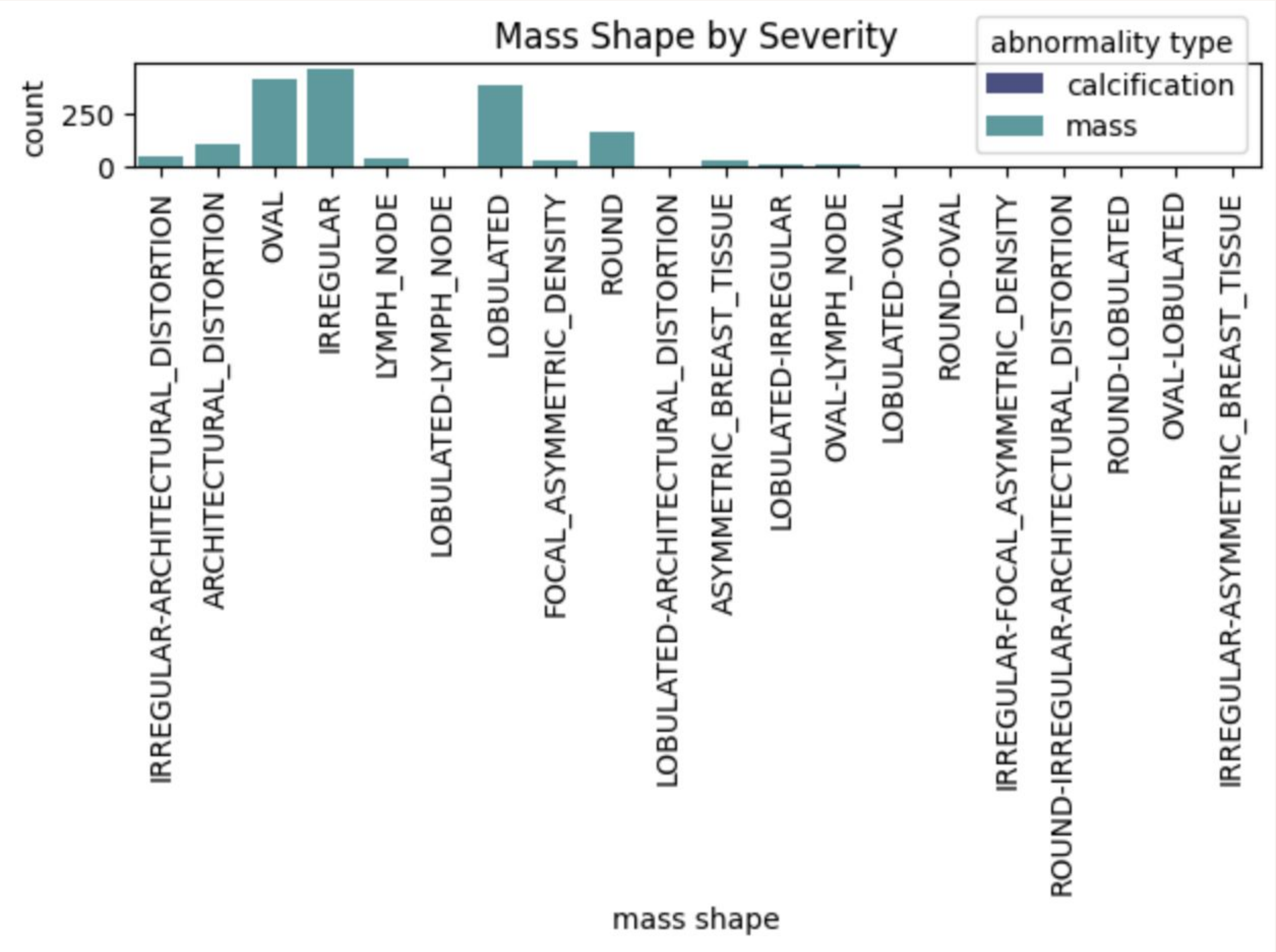
- **Loaded multiple CSVs** for mass and calcification case descriptions from train and test datasets.
- **Merged datasets** and identified key columns such as mass shape, assessment, pathology, abnormality type, and breast density.
- Handled **missing values** across columns, particularly in breast density and subtlety.
- Standardized and **simplified rare or composite mass shape categories** (e.g., grouping hybrid labels like ROUND-OVAL, LOBULATED-IRREGULAR).
- Generated **descriptive statistics** to understand distributions of assessment, subtlety, and breast density.

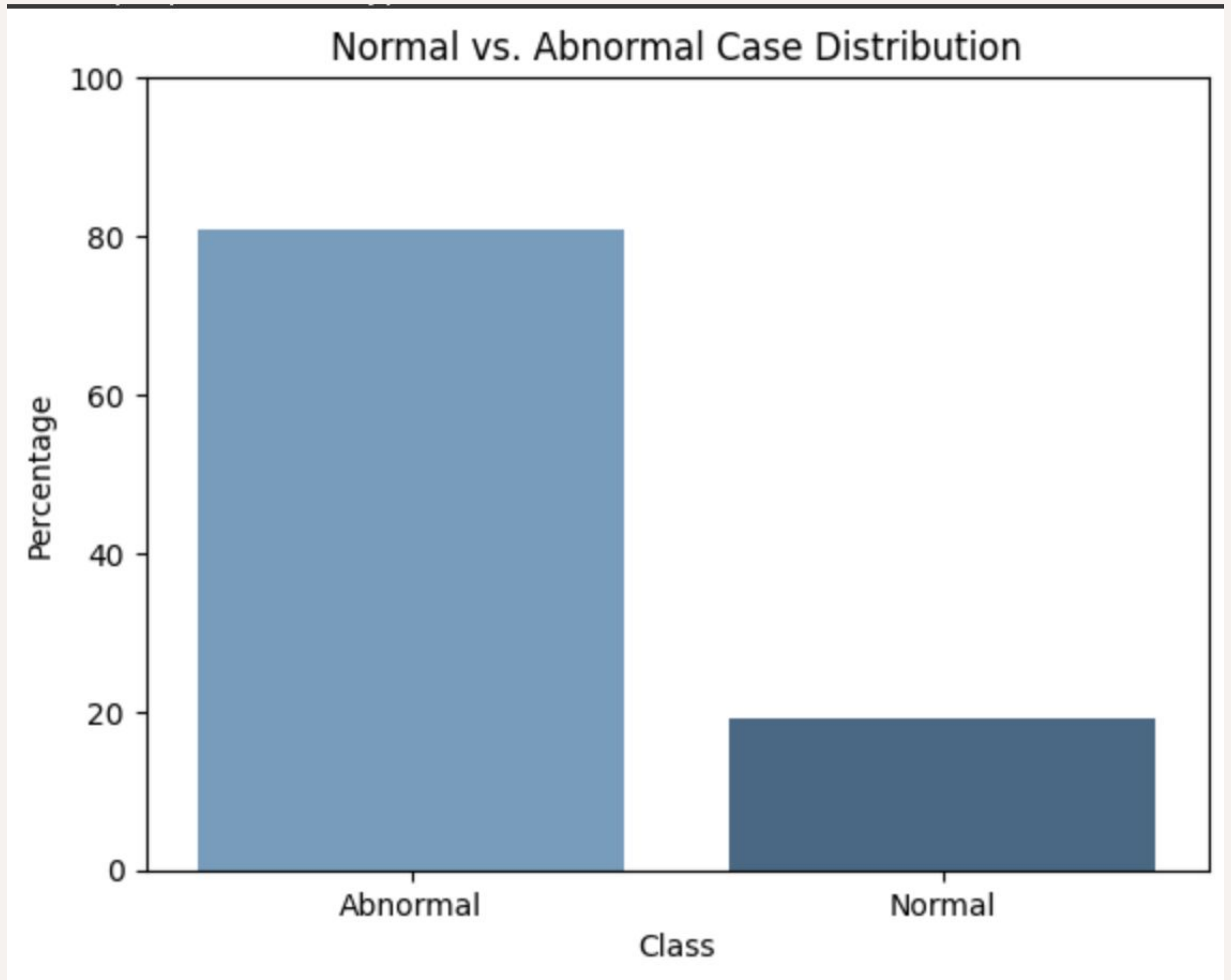
- Created **bar plots** to visualize:
  - Mass shape distribution.
  - Abnormal vs. normal case percentages (showing ~80% abnormal).
  - Relationships between mass shape and pathology (malignant, benign).

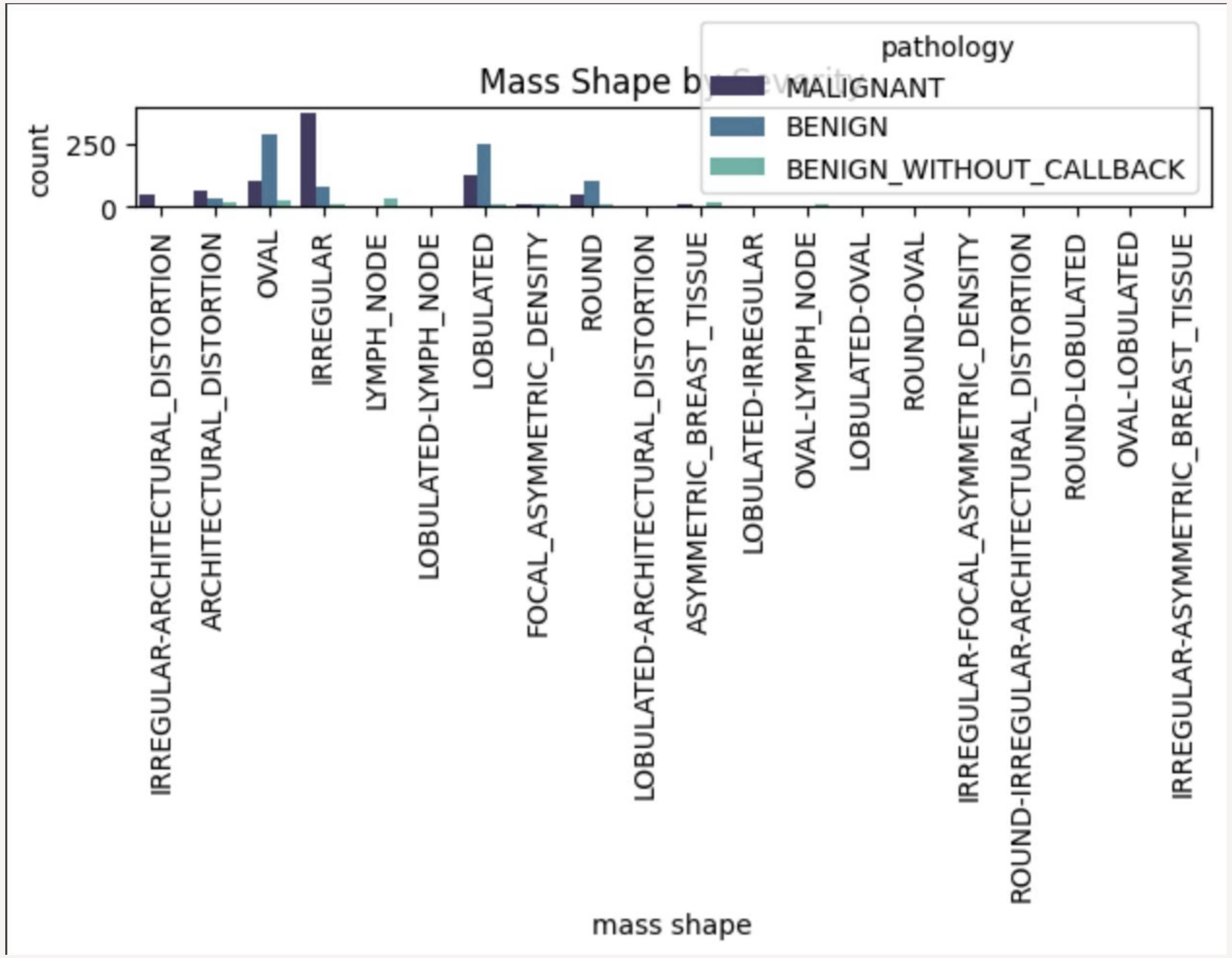
- These insights helped **identify dominant patterns**, such as IRREGULAR,

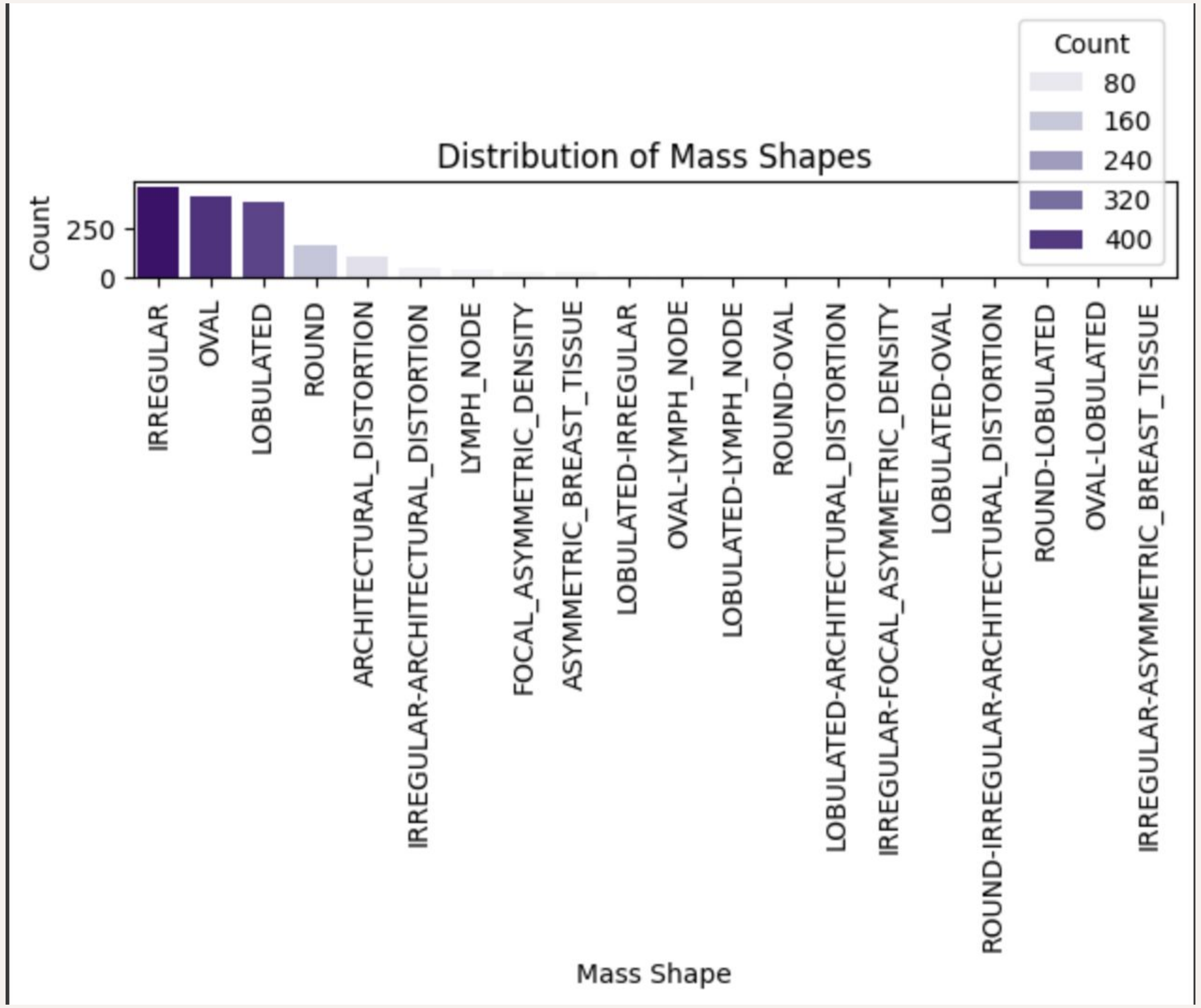
  OVAL,and LOBULATED being the most frequent mass shapes, with

  IRREGULAR showing higher malignancy rates.






Mass Shape by Severity
pathology
MALIGNANT
BENIGN
BENIGN_WITHOUT_CALLBACK
count
250
0
IRREGULAR-ARCHITECTURAL_DISTORTION
ARCHITECTURAL_DISTORTION
OVAL
IRREGULAR
LYMPH_NODE
LOBULATED-LYMPH_NODE
LOBULATED
FOCAL_ASYMMETRIC_DENSITY
ROUND
LOBULATED-ARCHITECTURAL_DISTORTION
ASYMMETRIC_BREAST_TISSUE
LOBULATED-IRREGULAR
OVAL-LYMPH_NODE
LOBULATED-OVAL
ROUND-OVAL
IRREGULAR-FOCAL_ASYMMETRIC_DENSITY
ROUND-IRREGULAR-ARCHITECTURAL_DISTORTION
ROUND-LOBULATED
OVAL-LOBULATED
IRREGULAR-ASYMMETRIC_BREAST_TISSUE
mass shape


Distribution of Mass Shapes
Count
80
160
240
320
400
250
0
IRREGULAR
OVAL
LOBULATED
ROUND
ARCHITECTURAL_DISTORTION
IRREGULAR-ARCHITECTURAL_DISTORTION
LYMPH_NODE
FOCAL_ASYMMETRIC_DENSITY
ASYMMETRIC_BREAST_TISSUE
LOBULATED-IRREGULAR
OVAL-LYMPH_NODE
LOBULATED-LYMPH_NODE
ROUND-OVAL
LOBULATED-ARCHITECTURAL_DISTORTION
IRREGULAR-FOCAL_ASYMMETRIC_DENSITY
LOBULATED-OVAL
ROUND-IRREGULAR-ARCHITECTURAL_DISTORTION
ROUND-LOBULATED
OVAL-LOBULATED
IRREGULAR-ASYMMETRIC_BREAST_TISSUE
Mass Shape

```
mass shape
IRREGULAR                                    464
OVAL                                         412
LOBULATED                                    384
ROUND                                        164
ARCHITECTURAL_DISTORTION                     103
IRREGULAR-ARCHITECTURAL_DISTORTION            52
LYMPH_NODE                                    35
FOCAL_ASYMMETRIC_DENSITY                      25
ASYMMETRIC_BREAST_TISSUE                      25
LOBULATED-IRREGULAR                            6
OVAL-LYMPH_NODE                                6
LOBULATED-LYMPH_NODE                           4
ROUND-OVAL                                     3
LOBULATED-ARCHITECTURAL_DISTORTION             2
IRREGULAR-FOCAL_ASYMMETRIC_DENSITY             2
LOBULATED-OVAL                                 1
ROUND-IRREGULAR-ARCHITECTURAL_DISTORTION       1
ROUND-LOBULATED                                1
OVAL-LOBULATED                                 1
IRREGULAR-ASYMMETRIC_BREAST_TISSUE             1
Name: count, dtype: int64
```

Assessment values (which may reflect severity or diagnostic confidence) have a **mean of 3.4** , indicating generally moderate to high suspicion levels in abnormal cases.

Subtlety scores , with a mean around 3.6, suggest that most abnormalities are **reasonably visible** , though there's a wide spread (0–5).

Breast density ranges from 1 to 4, with the median at 2 — indicating most patients fall into **scattered or heterogeneously dense tissue categories** .

A small number of missing values are present in breast_density, which may need imputation or filtering.

	breast density	abnormality id	assessment	subtlety	breast_density
count	1872.000000	3568.000000	3568.000000	3568.000000	1696.000000
mean	2.669338	1.252242	3.396581	3.647422	2.246462
std	0.932322	0.705416	1.314327	1.182583	0.874071
min	0.000000	1.000000	0.000000	0.000000	1.000000
25%	2.000000	1.000000	3.000000	3.000000	2.000000
50%	3.000000	1.000000	4.000000	4.000000	2.000000
75%	3.000000	1.000000	4.000000	5.000000	3.000000
max	4.000000	7.000000	5.000000	5.000000	4.000000

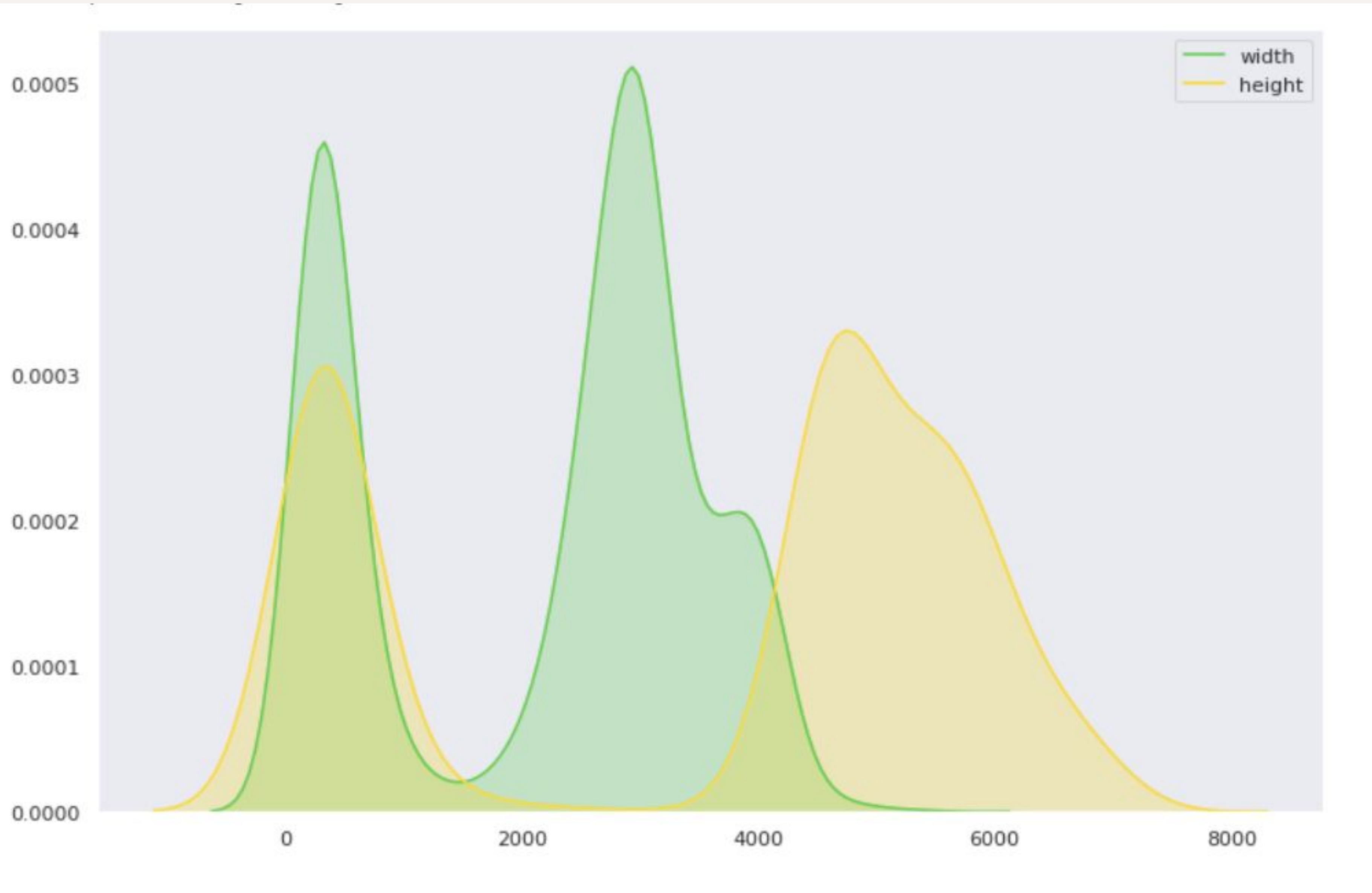
Image Types Analyzed : Dataset includes three major types:

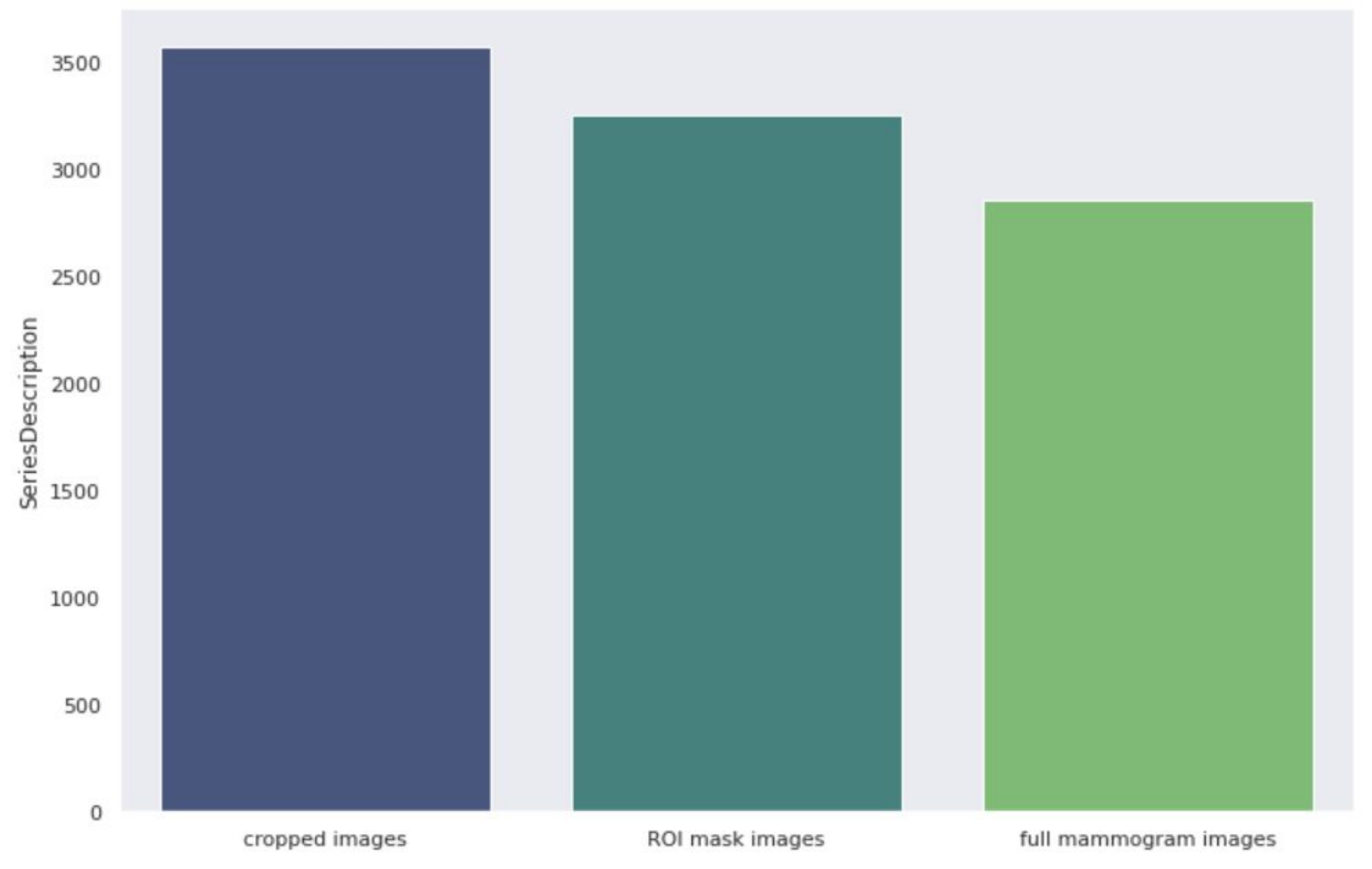
- Cropped images
- ROI (Region of Interest) mask images
- Full mammogram images

→ **Cropped images are the most common** , followed by ROI and full views.

Width and Height Distributions :

- Width and height values are **bimodally distributed** , suggesting distinct image types or resolutions.
- Majority of images cluster around two resolution zones.



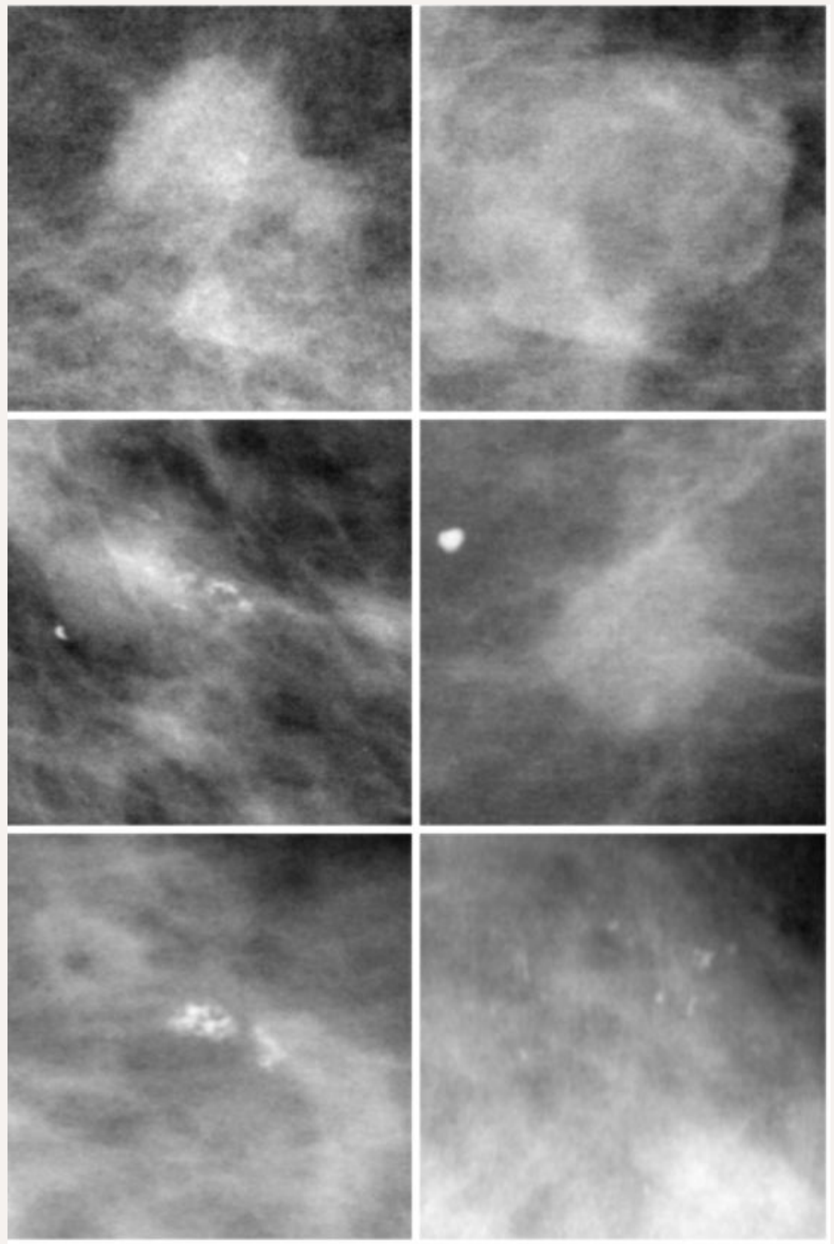


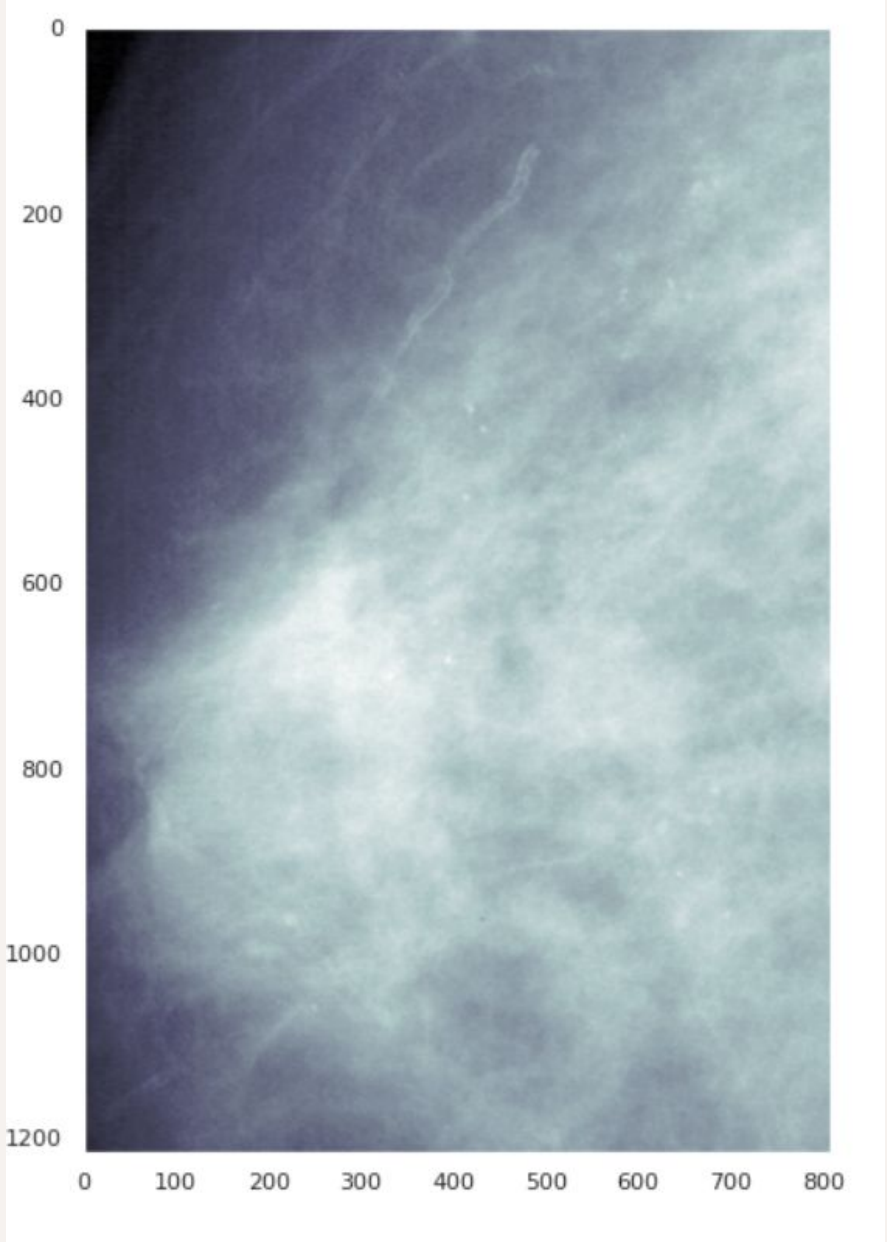
Aspect Ratio Insights :

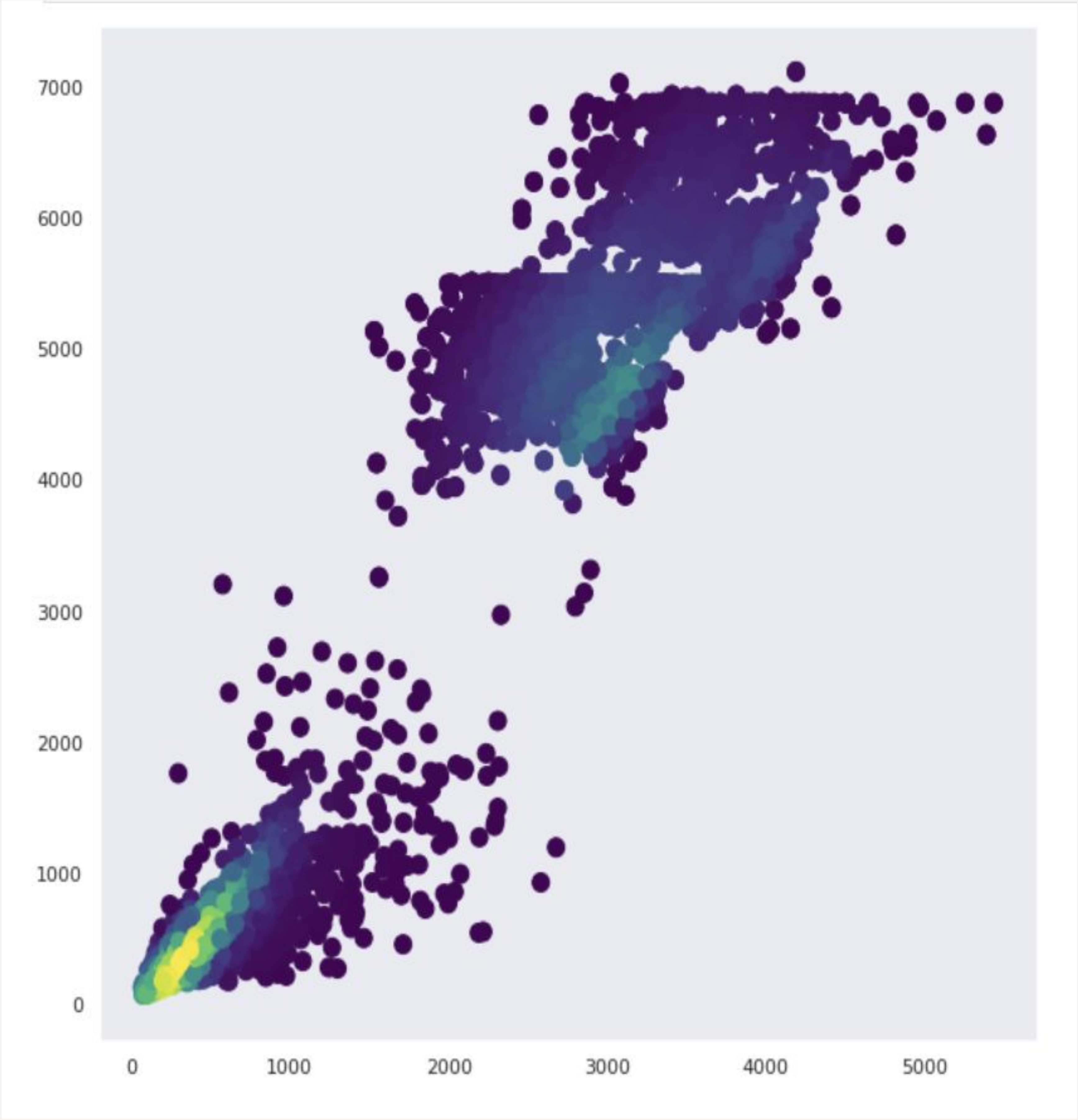
- Density scatter plot shows **distinct bands** in resolution clusters, indicating consistent acquisition settings for subsets of the data.
- Applied Gaussian KDE for enhanced visualization of pixel density regions.

Sample Image Visualization :

- Sampled and displayed grayscale mammogram to visually inspect image quality and tissue detail.







Preliminary Results

- **Mass Shape Distribution** :
  The majority of abnormal findings are associated with **irregular (464 cases)** , **oval (412)** , and **lobulated (384)** shapes. These categories dominate the dataset and may carry higher predictive value.
- **Assessment & Subtlety Scores** :
  The average **assessment score is ~3.4** and **subtlety ~3.6** , suggesting that most abnormalities are moderately to highly suspicious and visibly detectable.
- **Breast Density Patterns** :
  Most patients have breast densities between **2 and 3** , indicating tissue that is not extremely dense but may still obscure findings in some cases.
- **Rare & Composite Labels** :
  A significant number of **low-frequency hybrid mass shapes** were observed (e.g., "ROUND-OVAL"), which may require grouping to reduce noise in predictive modeling.

How Our Analysis Supports Modeling

Mass Shape Insights
→ Identified dominant mass shapes (irregular, oval, lobulated) to prioritize during model training and balance the dataset classes.

Label Cleaning and Standardization
→ Grouped rare and hybrid labels to **reduce label noise**, helping the model learn **clearer, more generalizable patterns**.

Assessment and Subtlety Scores
→ Can be incorporated as **auxiliary features** alongside image data to **enhance malignancy prediction** accuracy.

Breast Density Information
→ Understanding tissue density helps the model account for **image quality variations** and **potential misclassifications**.

Data Quality Improvements
→ Removing missing values and normalizing image sizes creates a **clean, high-quality input pipeline** crucial for fine-tuning the Vision-Language Model (VLM).

Strategic Prompt Design
→ Clinical labels (e.g., “A photo of an irregular mass with subtle features”) crafted from analysis can **strengthen text encoder input** during CLIP training.

Machine Learning Models Planned

- **Primary Approach:** Vision-Language Model based on CLIP (Contrastive Language–Image Pretraining).
- **Encoders Used:** Vision Transformer (ViT) for images and Transformer-based text encoder.
- **Purpose of Models:**
  - To fuse visual and textual clinical data for improved malignancy prediction.
  - To reduce dependence on manual radiologist interpretation by automating feature extraction.

Next Steps

Model Preparation

- Load pre-trained **CLIP** model components (image encoder and text encoder).
- Freeze encoders and design new **Fully Connected (FC) fusion layers** for domain-specific feature learning.

Training Pipeline Development

- Train the FC layers using mammogram-text pair similarities.
- Validate performance using metrics such as accuracy, precision, and recall.

Deployment Phase Setup

- Configure inference pipeline: **only use image encoder + trained FC layers** for malignancy prediction (benign vs. malignant).

GitHub Repository Overview

- Repository Link: github.com/AdityaDREXEL/CLIP_for_Breast_Cancer
- The repository is public and actively maintained by the team.
- The repository will be updated with model outputs, application interface code, and final documentation.

Thank You!

We are open to questions