Topic: Perform Exploratory Data Analysis Alice Tang and Yinda Chen

TF Lite Breast Cancer Detection Week 3: Perform EDA

1. Dataset Partition Strategy Review

For this project, we decided to use the pre-defined partitioning provided by the CBIS-DDSM: Breast Cancer Image Dataset, which was split by the dataset authors into different training and test sets. The dataset we are using consists of two primary subcategories— mass and calcification— with each one containing their own respective training and test splits. The split of the data is as follows:

“mass” Dataset

- Training Set: 71.32% (1,318 entries)
- Test Set: 28.68% (378 entries)

“calcification” Dataset

- Training Set: 21.02% (1,546 entries)
- Test Set: 78.98% (326 entries)

With careful deliberation, we decided to keep this partitioning instead of creating our own. However, there were several key factors that contributed to our decision:

1. **Dataset Author Intent**: We believe that the authors of this dataset created these splits with particular considerations in mind, ensuring that these splits will cultivate a robust training set for the model to learn from, all while keeping sufficient test data to evaluate actual performance. By keeping this data partition strategy, we establish that our results align with the original intent of the dataset, which in turn, will provide consistency and reliability in our outcomes.
2. **“pathology” Variable Distribution**: There is a bit of variation between the "mass" and "calcification" splits. This is due to the differences in the distributions of the target variable, "pathology" (malignant, benign, benign-without-callback). Each dataset contains different characteristics of the images, and because the underlying conditions vary between masses and calcifications, the customized splits aim to capture the unique patterns relevant to each category. The predetermined ratios will ensure that the model can recognize a balanced variety of examples for both types of abnormalities.
3. **Maximizing Training Data**: The training sets, which make up a larger portion of the dataset, provide the model with ample examples to learn from. This is paramount in medical imaging, as each patient is unique, for training a robust model capable of detecting easy-to-miss patterns in

the images. The splits for the "mass" and "calcification" datasets— approximately 70-30 and 80-20, respectively—are regularly used in machine learning to strike a balance between providing sufficient training data while mitigating the risk of overfitting, and thus encouraging better generalization.

4. **Efficiency and Focus on Model Development**: Utilizing the pre-split datasets splits is also a means of efficiency. We find it more beneficial to focus our efforts on model development, feature extraction, and evaluation, rather than manually re-splitting the dataset. The current partitioning is well-structured and likely intentional, making further adjustments unnecessary at this time. If needed, cross-validation techniques will be employed in the future to fine-tune any performance metrics.

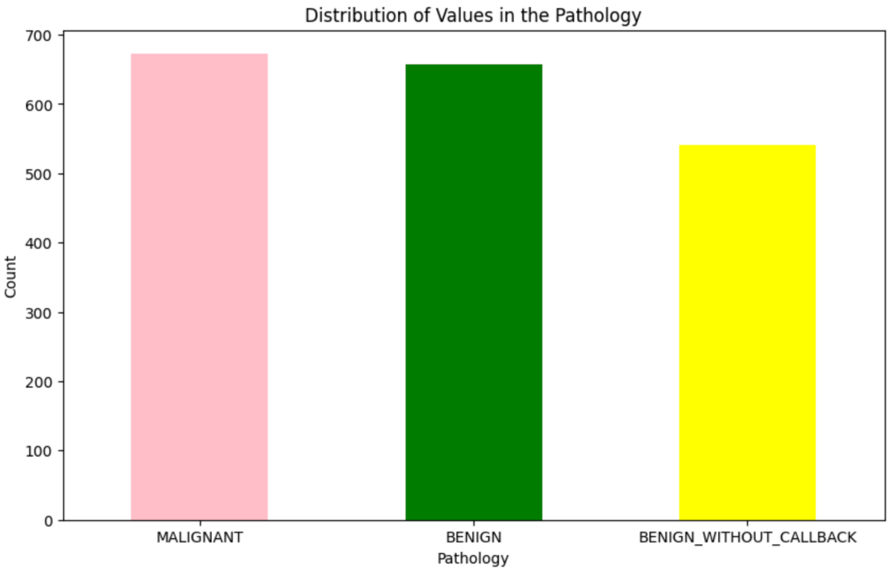
Ultimately, we decided to stand by the original dataset partitioning strategy, as it reflects the thoughtful design by dataset creators. This approach maximizes the model's exposure to the diverse data during training while promoting a reliable evaluation process, ensuring consistency and reproducibility in the results.

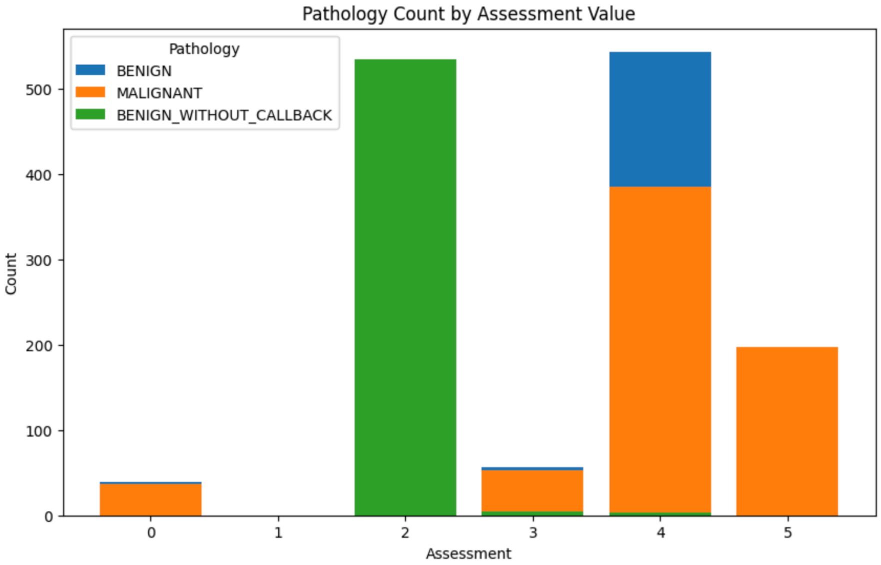
2. EDA Visualizations and Overview

In the initial EDA, we focused on key aspects of the dataset, particularly identifying missing data patterns such as mass_margins and calc type. We also conducted a thorough distribution analysis of pathology classes, breast density, and mass shapes, alongside a correlation analysis to explore the relationships between breast density and abnormality types.

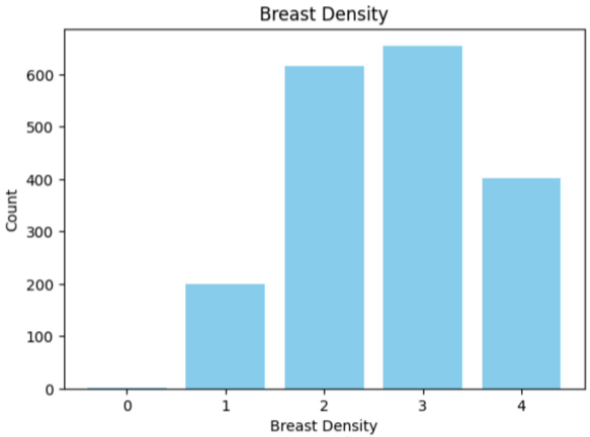
The figures below provide a visual summary of key insights gained, with more in-depth explanations discussed in the next sections.

I. Calcification Dataset EDA

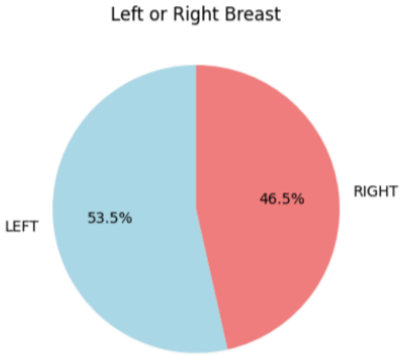




Pathology Count by Assessment Value



Breast Density



Left or Right Breast

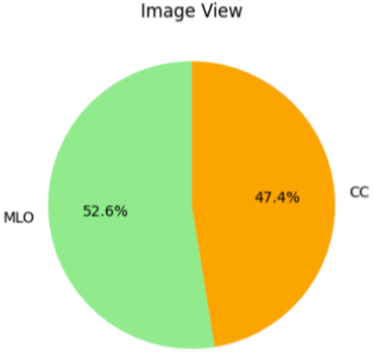
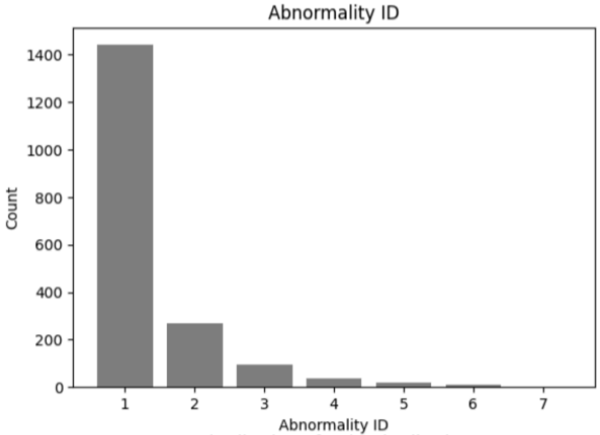
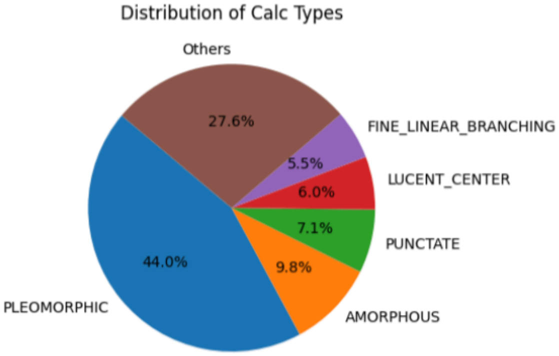


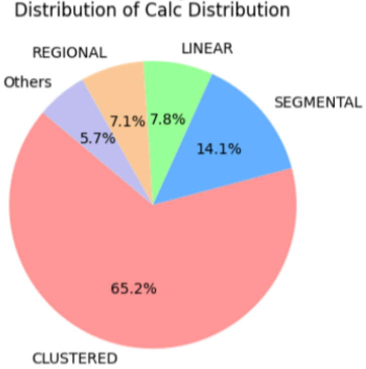
Image View



Abnormality ID

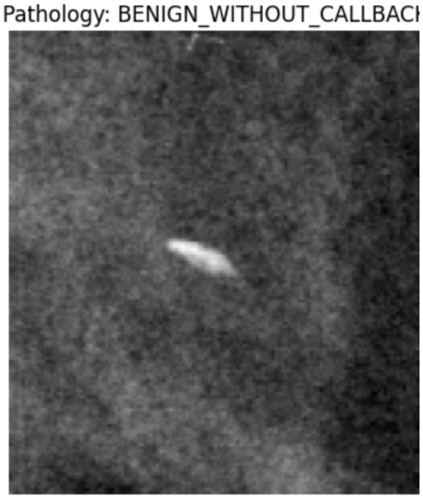


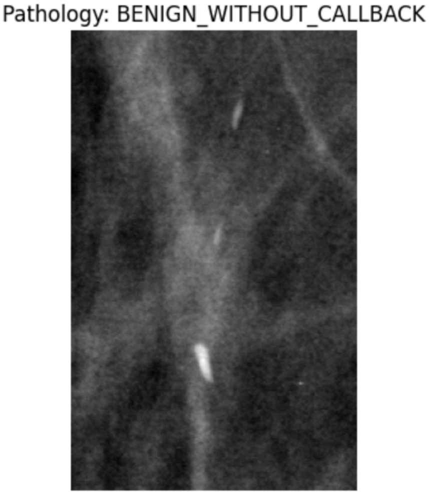
Distribution of Calc Types

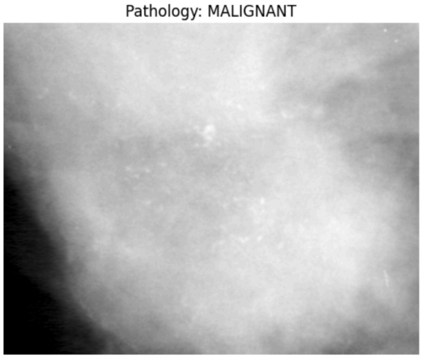


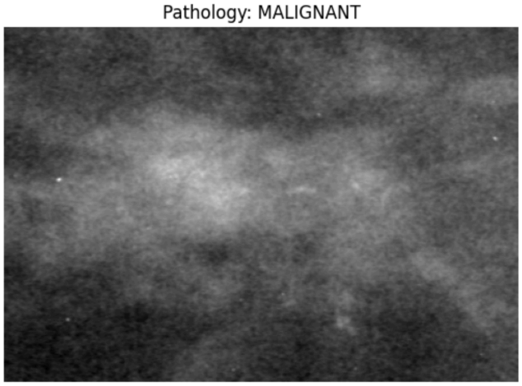
Distribution of Calc Distribution

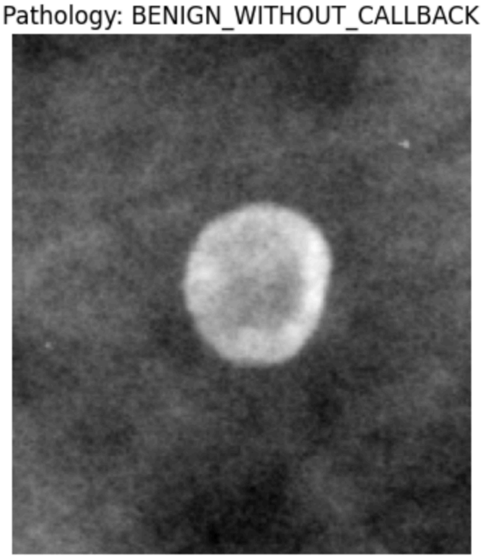
Calcification Example Images:



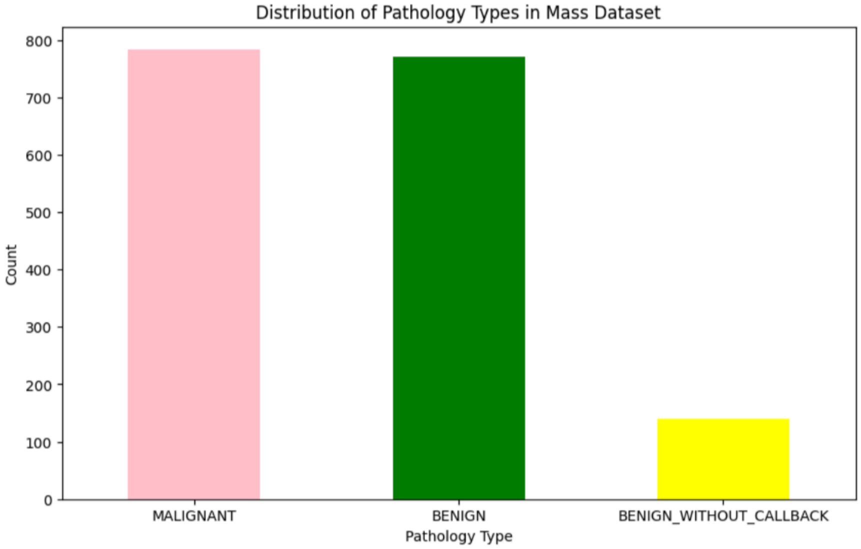


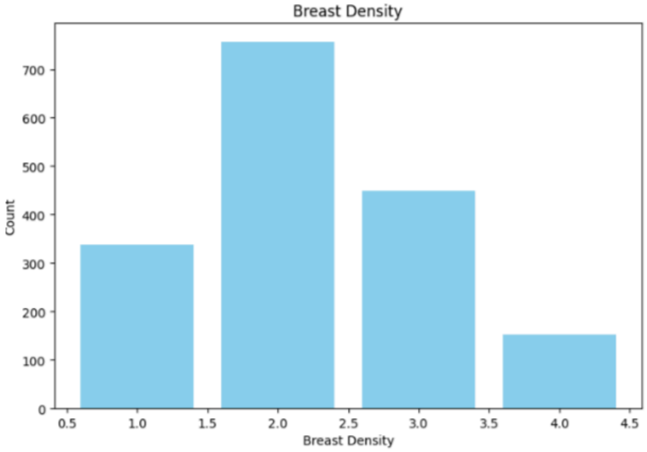


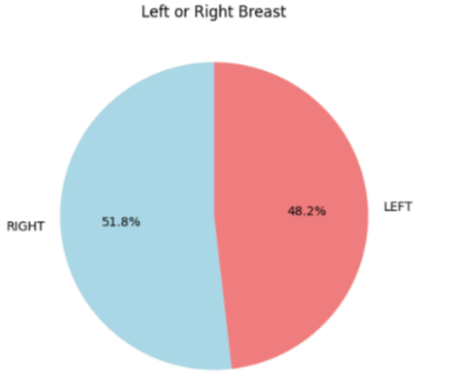


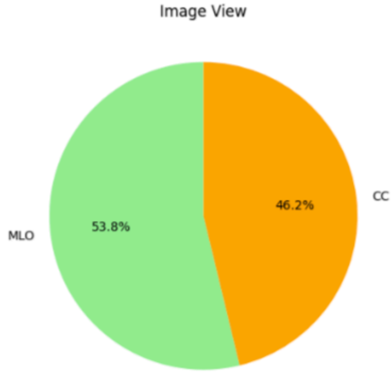


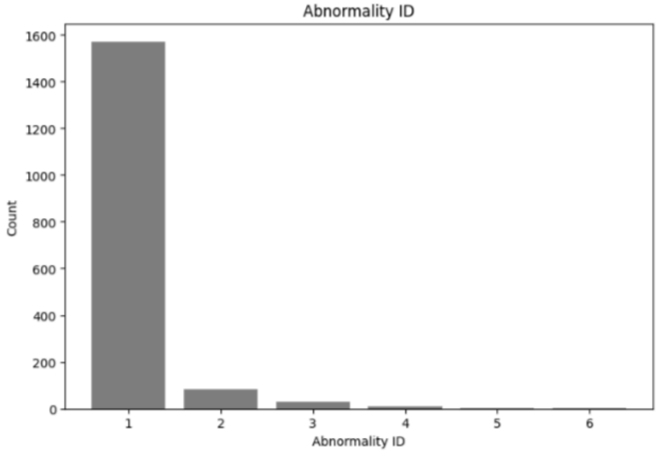
II. Mass Dataset EDA

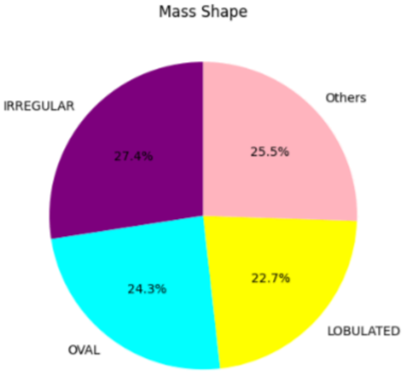


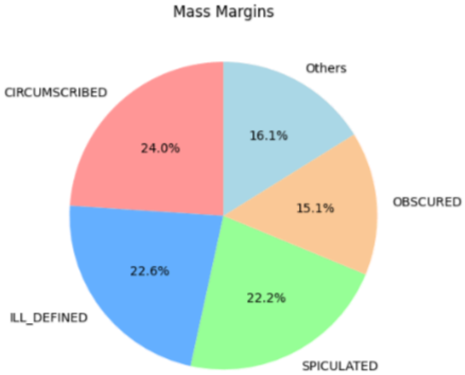




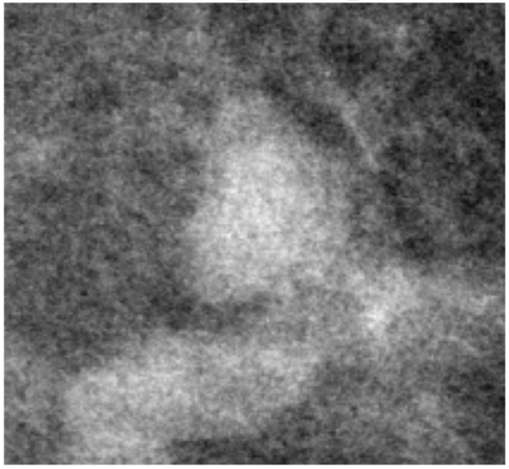




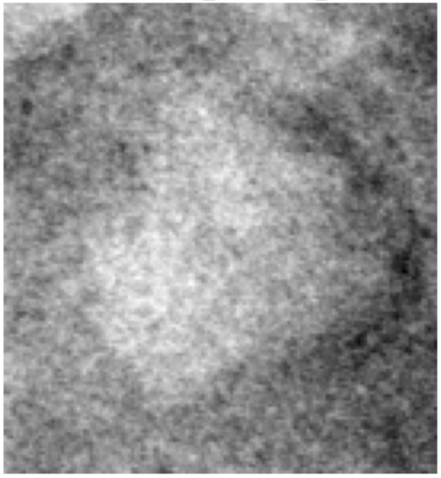




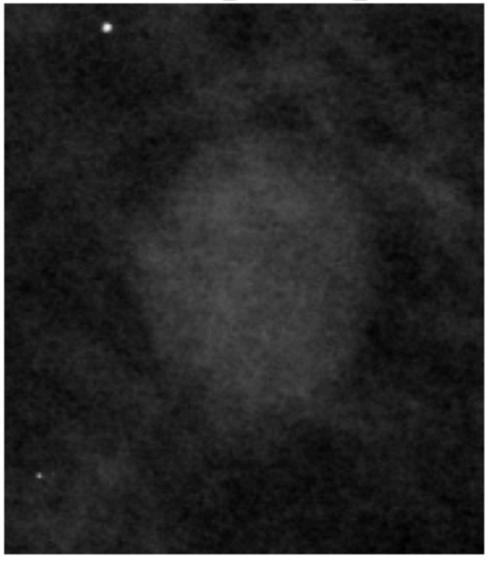
Mass Example Images:



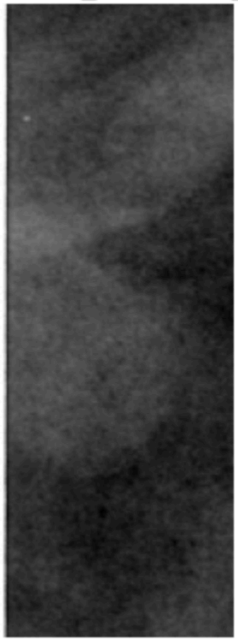
Pathology: BENIGN_WITHOUT_CALLBACK



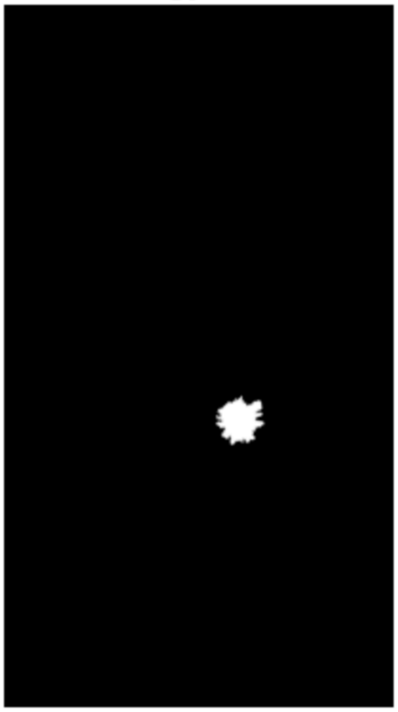
Pathology: BENIGN_WITHOUT_CALLBACK



Pathology: BENIGN_WITHOUT_CALLBACK



Pathology: BENIGN_WITHOUT_CALLBACK



Pathology: BENIGN

2.1 Insights and Challenges from Exploratory Data Analysis

The exploratory data analysis (EDA) of the CBIS-DDSM dataset uncovered multiple key insights that directly inform our problem statement, which is centered around the detection and classification of breast cancer using mammography images. The dataset is diverse, but that also implies that there are specific challenges and different opportunities that will impact the overall modeling approach.

1. **Handling Missing Data**: One of the biggest challenges for these datasets will be handling the missing values. The missing values are apparent within essential variables for characterizing breast abnormalities such as **calc_type**, **calc_distribution**, **mass_margins**, and **mass_shape**. The lack of clarity of these variables may impact the model's ability to learn from the data effectively— especially when it comes to calcification patterns and mass characteristics. Addressing these missing values will require careful consideration, whether through imputation or model adjustments, to preserve data integrity and optimize predictive performance.
2. **Balanced Pathology Classes**: A key strength of this dataset is its well-balanced distribution of pathology classes. There is a similar number of malignant (1,457), benign (1,429), and benign-without-callback (682) cases. Having a dataset with already balanced classes is a huge advantage, since it reduces the risk of class imbalance bias. This distribution provides the model with the means to be able to generalize across different pathologies, which is vital for clinical applicability.
3. **Breast Density and Imaging Views**: The dataset is primarily comprised of moderate (density 2) and high (density 3) breast density, with less images in the most and least dense categories. Due to the limited data in these categories, this may imply that the model may not be able to generalize well in rarer cases. Furthermore, the MLO (mediolateral oblique) view is more frequently used than the CC (cranio-caudal) view, which may also introduce some bias if this is not taken into consideration in model development. This will be a major problem, and we plan on understanding and addressing these biases to ensure the model still performs well across different clinical scenarios.
4. **Opportunities in Mass and Calcification Features**: The EDA highlighted the key characteristics of breast abnormalities, like the dominant calcification types (e.g., pleomorphic) and common mass shapes (e.g., irregular). These features are very promising in providing strong predictive potential for distinguishing between benign and malignant cases. For example, pleomorphic calcification types are frequently linked to malignant cases, while irregular mass shapes lean toward malignancy compared to round or oval forms. However, the variability in both mass shapes and margins will be a large obstacle to overcome, as the model must accurately capture these subtle differences to enhance diagnostic precision.
5. **Comparison of "calc_case" Images With "mass_case":** We noticed that the calc_case images do not exhibit any significant differences compared to the mass_case images. Moving forward, we will have to explore strategies in incorporating all three types of images— calc_case, mass_case, and any additional categories— into our analysis. The goal is to enhance the model's robustness and accuracy by utilizing the unique features of each image type. We will also need to investigate potential preprocessing techniques and feature extraction methods to optimize our results.

From the data exploration, it is clear that there are both challenges and opportunities within the CBIS-DDSM dataset. The dataset's well-balanced pathology distribution, diversity of imaging features, and detailed abnormality descriptors will be a solid foundation for model development. Notably, there will also be challenges with missing data and the potential for limited generalization due to imbalances in breast density and imaging views— all of which will need to be taken in account for.

2.2 Data Issues Identified and Recommendations for Pre-Processing

We have already highlighted a few key data issues within the CBIS-DDSM dataset, but we will now outline it in more detail:

1. **Missing Values:**
   - In the calcification datasets (calc_case_test and calc_case_train), important values like calc_type and calc_distribution contain missing values. In particular, calc_type has 322 non-null entries in the test set and 1,546 in the training set, while calc_distribution contains 263 non-null entries in the test set.
   - Similarly, in the mass datasets (mass_case_test and mass_case_train), the columns mass_shape and mass_margins are also incomplete, with mass_shape having 1,314 non-null entries in the training set, and mass_margins showing similar inconsistencies.
2. **Imbalance in Missing Data:**
   - Though the majority of columns are complete with data, some of the most important predictor variables like calc_type, mass_shape, and mass_margins have various levels of missing values. If not addressed, these data inconsistencies may skew results or introduce bias during model training.
3. **Data Quality:**
   - Specifically in the dicom_info dataset, multiple columns related to patient and study metadata (e.g., AccessionNumber, PatientBirthDate, PatientSex) are void of any usable information and may need to be removed from the analysis. Furthermore, the variability in breast abnormality attributes, such as mass shapes and margins, will be an obstacle in ensuring consistent and reliable inputs for the model.

Recommendations for Data Pre-Processing:

1. **Imputation for Missing Values:**
   - *Calcification and Mass Datasets:* To handle missing values in columns like calc_type, calc_distribution, mass_shape, and mass_margins, we would recommend employing imputation strategies. Possible approaches include filling missing values with the most common entries or domain-specific insights to make more informed imputations.

    - For columns with a higher amount of missing values, imputation may not be appropriate. We will consider excluding these entries to mitigate the risk of data skew.

2. **Feature Selection and Data Cleaning:**
    - *dicom_info Dataset:* For columns with minimal data such as AccessionNumber, PatientBirthDate, and PatientSex, we plan to exclude them from the analysis. These columns lack sufficient information and cannot be supplemented with publicly available data.
    - *Image Attributes:* We will ensure that the most critical variables such as breast_density, abnormality_id, and image view types are pre-processed correctly, as these factors will be a key player in image-based cancer detection models.
3. **Addressing Variability in Predictive Features:**
    - The dataset encompasses various mass shapes (e.g., irregular, round, oval) and mass margins (e.g., circumscribed, spiculated, obscured), which are essential for predicting malignancy. Proper pre-processing of these features, including potentially using dimensionality reduction or advanced feature engineering techniques) will be necessary to ensure the model effectively captures these subtle variations and enhances its predictive accuracy.

Through effectively addressing these identified data problems and applying the recommended pre-processing steps, the dataset can be utilized to its full potential for robust and accurate breast cancer detection modeling. This, in turn, will ensure that the data used is both comprehensive and suitable for predictive analysis.