Topic: Dataset Ingestion and Exploration Alice Tang and Yinda Chen

TF Lite Breast Cancer Detection Week 2: Dataset Ingestion and Exploration

1.1 Introduction

This week marks a pivotal point in our project as we transition to the **CBIS-DDSM: Breast Cancer Image Dataset**. We chose to switch to this dataset because it provides a more holistic representation, with a richer set of predictor variables, higher-quality image data, and multi-class labels that allow for more detailed classification (benign, benign-without-callback, malignant). Although our initial work began with the DDSM Mammography dataset, which also supports multi-class labels, we've decided to treat it as supplementary data moving forward. Given that it's still early in the project (Week 2), this is an ideal time to make this change, ensuring our analysis will benefit from the improved features and broader data available in CBIS-DDSM. Our main objective is still to develop a model capable of accurately classifying mammograms as positive or negative for cancer and to explore multi-class classification to enhance our understanding of varying cancer types. The success of this project will depend heavily on effective pre-processing and feature extraction, key steps in transforming raw image data into meaningful inputs for the model.

1.2 New Dataset Description

The CBIS-DDSM dataset is an updated and standardized version of the Digital Database for Screening Mammography (DDSM). Originally, the DDSM contained 2,620 scanned film mammography studies, covering normal, benign, and malignant cases, all verified with pathology information. The dataset's extensive nature and detailed ground truth annotations make it an invaluable resource for developing decision support systems in mammography.

CBIS-DDSM enhances the original dataset by providing decompressed images converted to DICOM format, along with updated ROI segmentation and bounding boxes. The dataset also holds detailed pathological diagnosis information, improving its usability in training and testing ML models. This curated subset, carefully selected by trained mammographers, certifies higher image quality and improved consistency for research— aligning with our project's needs.

We chose the CBIS-DDSM dataset because it addresses a key challenge in mammography research—the lack of a standardized, publicly accessible evaluation dataset. By using CBIS-DDSM, we are aligning with a well-curated resource that allows for more reproducible results and offers a robust foundation for developing and evaluating our model. The larger set of high-quality labeled images and the clear diagnostic annotations make it a superior option for our project compared to the DDSM dataset alone.

2. Selecting the Target Variable

The target variable for this project is labeled as “pathology", which is present across multiple datasets including testing and training for **calc_case_description**, **mass_case_description**, and others. The **"pathology"** column categorizes mammogram results into three different classes: **BENIGN**, **BENIGN_WITHOUT_CALLBACK**, and **MALIGNANT**. This multi-class classification allows the model to provide a more detailed diagnostic prediction. This is crucial in distinguishing not only between cancerous and non-cancerous findings but also between benign cases that may not require follow-up.

We selected this target variable because it aligns with the primary goal of our project: to develop a model capable of supporting medical professionals by identifying the type and severity of abnormalities in mammography images. While binary classification (e.g., breast cancer vs. no breast cancer) could have been another viable option, we believe that the inclusion of additional classes for benign cases offers a more holistic approach— potentially leading to more clinically useful outcomes.

Moreover, the multi-class nature of the "pathology" column provides a more detailed understanding of breast cancer cases. It mirrors real-world medical scenarios where the differentiation between types of benign abnormalities, such as calcifications and masses, can influence treatment recommendations and patient management strategies. This distinction is crucial in clinical practice, as it helps healthcare providers minimize unnecessary interventions while still identifying cases that may need additional monitoring. This approach not only improves patient care and experience but also helps to lower overall healthcare costs. Consequently, the "pathology" column is an ideal choice as the target variable, supporting the development of a sophisticated, real-world applicable decision support system.

3. Predictor Variables

In developing our model for breast cancer detection, we are utilizing four distinct datasets (not including the test datasets for calc_case and mass_case): **calc_case_description, mass_case_description, dicom_info,** and **meta_info**. These predictor variables were selected due to their relevance to mammographic assessments, and their potential to enhance the model’s accuracy and interpretability.

3.1 Predictor Variables from calc_case_description and mass_case_description

- **Breast Density:** This variable represents the density of breast tissue, which is crucial in mammographic evaluations. This is a paramount predictor variable, since dense tissue often obscures tumors, making it essential for accurately identifying potential abnormalities.
- **Left or Right Breast:** Determining which side the imaging belongs to will help in understanding the context of abnormalities and evaluating asymmetries. This variable will be important in

providing comprehensive diagnostic information and monitoring abnormalities over time, which can aid in early detection and patient management.

- **Image View:** The orientation of the mammogram (e.g., cranio-caudal, mediolateral oblique) will highly affect the interpretation of images. Because different orientations may highlight or obscure specific abnormalities, this variable will be vital for comprehensive analysis.
- **Calcification Type and Distribution (for calc_case_description):** The type and distribution of calcifications are essential for diagnosing specific conditions. These variables were chosen because they offer detailed information that helps differentiate between benign and malignant calcifications.
- **Mass Shape and Margins (for mass_case_description):** The shape and margins of the detected masses aid in assessing the probability of being benign or malignant. This variable will be crucial for distinguishing between the different types of masses.
- **Assessment:** This overall evaluation of the mammogram provides an overarching judgment that will reflect in the likelihood of malignancy. It acts as a key predictor for determining the overall risk associated with the findings.
- **Subtlety:** This measures how noticeable an abnormality is in an image. Higher values of subtlety correspond to more pronounced abnormalities. This measure will be vital for precise detection and diagnosis because visible abnormalities are easier to identify.

By integrating these predictor variables, we aim to build a model that not only captures detailed and relevant information from mammographic images but also mirrors real-world clinical practices. In addition to the key variables identified, we will utilize other valuable information available in the dataset to further enhance the model's predictive power. Furthermore, we can analyze the pixel-level data from each image to directly extract critical details, or alternatively, leverage a pre-trained model (such as VGG-16 or ResNet) to automatically extract features, a technique that has proven highly effective in previous cancer detection tasks[1]. This multifaceted approach will provide deeper insights, helping the model deliver precise and actionable results. In tandem, this comprehensive strategy will facilitate an environment for more accurate breast cancer detection, assist in risk stratification, and aid in effective patient management.

[1] Bakasa, Wilson, and Serestina Viriri. "VGG16 Feature Extractor With Extreme Gradient Boost Classifier for Pancreas Cancer Prediction." *Journal of Imaging*, vol. 9, no. 7, July 2023, p. 138. https://doi.org/10.3390/jimaging9070138.

4. Exploration of Dataset

The CBIS-DDSM dataset is comprised of several key components, each providing different characteristics relevant to breast cancer detection.

4.1 calc_case_test and calc_case_train Dataset Information

The calc_case_test data frame includes 326 rows and 14 columns, with a mix of data types with four integer columns and ten string/object columns. This dataset primarily provides information about patient IDs, breast density, breast side, image views, abnormalities, assessments, and file paths related to images and masks. The columns calc_type and calc_distribution contain missing values, with 322 and 263 non-null entries respectively, while the remaining columns are complete with information. The calc_case_train data frame mirrors the calc_case_test in structure but with a larger dataset of 1,546 entries. Missing values are similarly present in calc_type and calc_distribution, following the same pattern as in the test dataset.

4.2 dicom_info Dataset Information

The dicom_info data frame is the largest and most extensive, with 10,237 rows and 38 columns encompassing a range of data types including integers, floating-point numbers, and objects. This dataset details medical imaging files, including file paths, image metadata, and patient information. However, certain columns such as AccessionNumber, PatientBirthDate, and PatientSex contain no valid data, whereas columns like file_path, image_path, and BodyPartExamined contain all information. Additionally, there are missing values present in the columns related to patient and study details.

4.3 mass_case_test and mass_case_train Dataset Information

In the mass_case_description_test_set, there are 378 rows and 14 columns. The dataset is complete except for mass_margins, which has some missing values with only 361 non-null entries. This dataset provides information about breast cancer cases, including patient ID, breast density, abnormality details, and image file paths.

Similarly, the mass_case_description_train_set has 1,318 entries and 14 columns. It features complete data except for mass_shape and mass_margins, which have some missing values with 1,314 and 1,275 non-null entries respectively. This training dataset provides the same information as the test set.

4.4 meta_info Dataset Information

The meta_info DataFrame contains 6,775 rows and 9 columns, with a mix of integer and object data types. Each row in this dataset offers details about medical imaging series, including SeriesInstanceUID, StudyInstanceUID, Modality, and SeriesDescription. Most notably, there are no missing values in this dataset.

4.5 Overall Statistics

In terms of overall statistics, the dataset is balanced between the different pathology types, with a slightly higher number of malignant cases. Specifically, there are 1,872 images related to calcifications and 1,696 images related to masses. Among these, 1,457 images are classified as malignant, 1,429 as benign, and 682 as benign without callback. This is a good sign, as having balance is paramount in training robust models that are capable of distinguishing between various pathology types effectively.

4.6 Further Exploration

Delving deeper into the exploration of the dataset reveals multiple new insights in understanding the data. For example, breast density is most commonly categorized as moderate (density 2) and high (density 3), with fewer images representing the densest (density 1) and sparsest categories (density 4). The distribution of images between the left and right breasts is fairly balanced, with only a slight preference for the left. Similarly, the MLO view is more frequently used than the CC view, reflecting its broader breast tissue coverage.

The abnormality ID distribution implies that most abnormalities are classified under ID 1, with fewer occurrences for other IDs. Calcification types and distributions are dominantly pleomorphic types and clustered distributions, while mass shapes have a variety of forms with irregular shapes being the most common. Mass margins also display a range of characteristics, with circumscribed margins being the most prevalent.

4.7 Summary

Overall, the exploration of the dataset shows that this is a well-structured and diverse collection of mammography data essential for breast cancer detection. With the various data types, variable definitions, and distribution patterns, this sets the stage for a more in-depth analysis of the data. Moreover, the balanced pathology categories and various imaging attributes facilitate an environment for robust model development and analysis. This week's assignment lays a solid foundation for future work and sets the scene for effective modeling. Understanding the data and its story is crucial for creating a strong model and ensures that our future analysis will utilize the dataset's full potential for accurate breast cancer detection.