Phase-1: Data Loading and Structuring

This phase establishes the foundation for working with the VinDr-Mammo dataset by designing a reproducible and verifiable pipeline for data ingestion, validation, and restructuring. The outputs of this phase ensure consistency across image data and metadata, enabling reliable downstream analysis and modeling.

---

1. Data Ingestion and Initial Preprocessing

Inputs

- Dataset ZIP archives and extracted folders.
- Metadata CSVs containing study information, bounding box annotations, findings, breast density, and BI-RADS ratings.

Operations

- Defined dataset root paths (`BASE_FOLDER`, `EXTRACT_PATH`, `CLEANED_PATH`).
- Verified file integrity and completeness.
- Standardized reusable paths for metadata and image storage (`FINAL_DF_PATH`, `STUDY_DF_PATH`, `IMAGE_DF_PATH`, `IMAGE_DF_SAVE_PATH`).

---

2. Image Organization

Objective
Align image storage with study identifiers to create a consistent folder structure.

Steps

- Grouped mammogram images by study ID.

- Segregated images into `NORMAL_FOLDER` and `ABNORMAL_FOLDER` using study labels.
- Performed consistency checks between expected and actual image counts (`expected_count`, `actual_count`).

---

3. Exploratory Data Analysis (EDA)

Metadata Profiling

- Age distributions per study (`age_median_per_study`).
- BI-RADS distribution analysis (`birads_counts`, `birads_binary_counts`, `birads_median_age`).
- Breast density category statistics (`density_counts`).
- Bounding box frequency counts (`bbox_count`).

Study-Level Exploration

- Number of findings per study (`all_findings`).
- Verification of study coverage and completeness (`df_study_ids`, `df_studies_set`).

---

4. Data Quality Checks

Consistency Validation

- Cross-validation between metadata entries and image folders (`existing_studies`, `extra_in_folder`).
- Identification of mismatched or missing abnormal studies (`abnormal_studies`, `abnormal_errors`).

- Validation of annotation alignment with images (`bbox_errors`).

Sanity Checks

- Detection of missing or duplicate study IDs.
- Median age validation across BI-RADS and breast density categories.
- Verification of bounding box–study mapping integrity.

---

5. Folder Restructuring

- Established a standardized folder hierarchy under `CLEANED_PATH`.
- Ensured reproducibility by mapping raw dataset organization into a clean, consistent structure suitable for experiments.

Key Variables

Category	Variables
Path Management	`BASE_FOLDER`, `FOLDER_ROOT`, `IMAGE_ROOT`, `CLEANED_PATH`, `EXTRACT_PATH`, `FINAL_DF_PATH`, `STUDY_DF_PATH`, `IMAGE_DF_PATH`, `IMAGE_DF_SAVE_PATH`, `NORMAL_FOLDER`, `ABNORMAL_FOLDER`, `breast_anno_path`
DataFrames and Sets	`all_studies_df`, `df_study_ids`, `df_studies_set`, `all_findings`
Quality Control	`expected_views`, `expected_count`, `actual_count`, `existing_studies`, `extra_in_folder`, `abnormal_studies`, `errors`, `abnormal_errors`, `bbox_errors`
Statistical Analysis	`birads_counts`, `birads_binary_counts`, `birads_median_age`, `density_counts`, `age_median_per_study`, `bbox_count`

Outputs of Phase-1

- A cleaned and structured dataset with a standardized folder hierarchy.
- Integrity reports covering study counts, BI-RADS distribution, breast density categories, and bounding box statistics.
- Baseline descriptive statistics for exploratory analysis.
- Quality control reports highlighting inconsistencies, missing data, and annotation mismatches.