Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	LTVIP2024TMID24981
Project Title	Deep learning techniques for breast cancer prediction
Maximum Marks	2 Marks

Data Quality Report Template

A **Data Quality Report** is crucial for assessing the quality of the dataset used in a breast cancer prediction project, especially when using a **Convolutional Neural Network (CNN)**. It helps identify issues such as missing data, poor-quality images, or class imbalances that may negatively impact the performance of the model. Here's a detailed template you can follow for creating a data quality report.

Data Source	**Data Quality Issue**	**Severity**	**Resolution Plan**
Break His Dataset	Missing labels for some images	High	Conduct a manual review of the images and relabel them.
DDSM (Digital Database for Screening Mammography)	Class imbalance (more benign than malignant)	Medium	Use SMOTE or other oversampling techniques for the minority class.
In-house Hospital Images	Low-resolution images	High	Remove low-resolution images or enhance resolution if feasible.
Public Dataset (e.g., CAMELYON)	Noisy images with artifacts	Medium	Filter out noisy images; consider data cleaning techniques.

Private Institution Collection	Duplicate images present	Low	Identify duplicates and remove them from the dataset.
Publicly Available Datasets	Inconsistent naming conventions	Medium	Standardize naming conventions across all files.
Local Hospital Imaging	Blurry images	Medium	Review images; remove or enhance using deblurring techniques.
Multiclass Dataset	Mislabeling of samples	High	Cross-check labels with domain experts and relabel as necessary.
Annotation Files	Missing metadata for some images	Medium	Ensure all images have corresponding metadata; fill in gaps.
Mixed-Quality Sources	Unbalanced distribution of age groups	Low	Stratify sampling to ensure representation across age groups.