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 Collection Plan & Raw Data Sources Identification Template

Objectives

- Purpose of Data Collection:
    - Describe the overall objective of collecting the data (e.g., to build a CNN model for predicting breast cancer from medical images).
- Specific Goals:
    - Define specific goals, such as achieving a certain accuracy rate or validating against established datasets.

Data Collection Plan Template

Section	Description
Project Overview	Provide a brief overview of the machine learning project, outlining its goals and objectives. *Example*: The project aims to develop a Convolutional Neural Network (CNN) model for predicting breast cancer based on histopathology images. The primary objective is to accurately classify tumor

	samples into benign and malignant categories to assist in early diagnosis and treatment planning.
Data Collection Plan	Outline the sources from which data will be collected, including both public and private datasets. *Example*: Data will be collected from a combination of publicly available datasets, such as the Break His and DDSM, as well as private institutional data from local hospitals. Collaborations will be established with medical institutions to obtain additional images and associated metadata.
Raw Data Sources Identified	For the **Raw Data Sources Identified**, provide details for each source, focusing on their relevance to your project.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Break His Dataset	A dataset of histopathology images of breast tumors. Contains multiple magnifications and classifications.	Break His	JPG, PNG	~7,909 images	Open access
DDSM (Digital Database	Comprehensive collection of mammogram	DDSM	DICOM	~3,000 images	Open access

for Screening Mammography)	images, includes various cancer stages and normal cases.				
CAMELYON Dataset	Focused on cancer metastasis detection in lymph nodes, offering high-quality histopathological images.	CAMELYON	TIFF	~1,000 images	Open access
Kaggle Datasets	User-uploaded datasets related to breast cancer diagnosis and analysis available for machine learning.	Kaggle	Varies	Varies by dataset	Open access, account required.