Data Collection and Preprocessing Phase

Date	22 October 2024
Team ID	SWTID1727274979
Project Title	Deep learning techniques for breast cancer risk prediction
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

This Data Collection Plan and Raw Data Sources Identification Template is a structured framework used to guide the systematic collection, organization, and documentation of data necessary for achieving specific objectives. This document ensures consistency, accuracy, and efficiency in the data-gathering process.

Data Collection Plan Template

Section	Description
Project Overview	The project aims to utilize deep learning techniques to predict the risk of breast cancer using histopathology images. The objective is to build a robust model capable of classifying images into benign and malignant categories based on the presence of cancerous cells. By leveraging advanced convolutional neural networks (CNNs), the project seeks to improve the accuracy of early cancer detection.
Data Collection Plan	Data is collected from publicly available online sources, specifically the Breast Histopathology Images Dataset available on Kaggle. The dataset contains microscopic images of breast cancer tissue samples, annotated as benign (Non – cancerous) or malignant(Cancerous) . The data will be used for model training, validation, and testing in the classification task .
Raw Data Sources Identified	The primary raw data source for this project is the Breast Histopathology Images Dataset available on Kaggle. This dataset consists of 277,524 patch images of breast tissue samples , including both benign and malignant tissue samples. The images are labeled to facilitate classification tasks, making it suitable for

	training deep learning models aimed at predicting breast cancer risk. The data is publicly accessible and can be downloaded from the Breast Histopathology Images Dataset on Kaggle. The images are stored in PNG format, and the dataset has a size of approximately 2.7 GB .

Raw Data Sources Template

Dataset Information : For this project only one dataset i.e Breast Histopathalogy images Dataset is used which is publicly available on kaggle . This image dataset contains a total of 277,524 patch images of breast tissue samples including both benign and malignant cases .

Source Name	Description	Location/URL	Format	Size	Access Permissions
Dataset 1	The dataset consists of breast histopathology images, categorized as benign (0) or malignant(1) . It provides labeled images in the form of 0 and 1 used for classification tasks, suitable for training deep learning models to predict breast cancer risk.	**Dataset Link/URL :** **https://www.kaggle.com/paultimothymooney/breast-histopathology-images**	Images (PNG) Format	2.7 GB	Public