Data Collection and Preprocessing Phase

Date	17 JUNE 2025
Team ID	SWTID1749825524
Project Title	Deepfruitveg: Automated Fruit and Veg Identification
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Report:

About the Dataset

Context

This dataset encompasses images of various fruits and vegetables, providing a diverse collection for image recognition tasks. The included food items are:

- **Fruits**: Banana, Apple, Pear, Grapes, Orange, Kiwi, Watermelon, Pomegranate, Pineapple, Mango
- **Vegetables**: Cucumber, Carrot, Capsicum, Onion, Potato, Lemon, Tomato, Radish, Beetroot, Cabbage, Lettuce, Spinach, Soybean, Cauliflower, Bell Pepper, Chilli Pepper, Turnip, Corn, Sweetcorn, Sweet Potato, Paprika, Jalapeño, Ginger, Garlic, Peas, Eggplant

Content

The dataset is organized into three main folders:

- **Train**: Contains 100 images per category.
- **Test**: Contains 10 images per category.
- **Validation**: Contains 10 images per category.

Each of these folders is subdivided into specific folders for each type of fruit and vegetable, containing respective images.

Data Collection Plan:

Section	**Description**
Project Overview	To design and implement a deep learning-based system capable of accurately identifying various fruits and vegetables from images using Convolutional Neural Networks (CNNs), streamlining classification and quality control in agriculture, retail, and food processing industries.
Data Collection Plan	• Source image datasets from **Kaggle**. • Choose datasets with **well-labelled folders** for each fruit/vegetable class. • Ensure **diverse conditions** (lighting, angle, background) for better model generalization. • Remove **corrupt/misplaced images** and **standardize formats** (e.g., JPG, PNG).
Raw Data Sources Identified	The raw data sources for this project include datasets obtained from Kaggle, the popular platforms for data science competitions and repositories. The provided sample data represents a subset of the collected information, encompassing variables such as test ,train , validate

Raw Data Sources Report:

Source Name	**Description**	**Location/URL**	**Format**	**Size**	**Access Permissions**
Kaggle Dataset	The dataset used for **DeepFruitVeg** is an **image-based multi-class classification dataset** containing photos of various fruits and vegetables. Each image is categorized into a labeled folder based on its class (e.g., Apple/, Banana/, Tomato/).	https://www.kaggle.com/datasets/kritikseth/fruit-and-vegetable-image-recognition	jpg	2GB	Public