Data Collection and Preprocessing Phase

Date	19 June 2025
Team ID	SWTID1750170729
Project Title	**Deepfruitveg: Automated Fruit And Veg Identification**
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
Kaggle | Plant Type Dataset Used	1. Images not uniformly sized (some > 378x498) 2. Some corrupted or unreadable images 3. Class imbalance (some classes had more than required samples) 4. Because of Large Dataset its taking more time to train our Model	Moderate	1. Resize all images to a fixed dimension (224x224 or 260x200) 2. Use cv2.imread() with a check for None to skip corrupted files 3. Limit number of images per class using random.sample() for balance 4. All images were resized to **260×200** to match **EfficientNetB3 input requirements**. Used cv2.resize() and skipped unreadable images. EfficientNetB3's preprocess_input() was used for pixel normalization.