Data Collection and Preprocessing Phase

Date	5 Feb 2026
Student Name	Rounak Pratap Gajbar
Project Title	greenclassify: deep learning-based approach for vegetable image

Data Quality Report Template

Data Quality Report

Data Source: Kaggle Vegetable Image Dataset (Primary Source) + Supplementary Images (if applicable)

Data Source	Data Quality Issue	Severity	Resolution Plan
Kaggle Dataset	Potential Class Imbalance	Moderate	Analyze class distribution. If significant imbalance is detected, apply data augmentation techniques (e.g., image rotation, flipping, brightness adjustments) to under-represented classes in the training set.
Kaggle Dataset	Presence of Low-Quality Images (blurry, poor lighting)	Low	Manually review a sample of images; if deemed necessary, remove images with severe quality issues from the training set. Automated quality filtering may be considered.
Kaggle Dataset	Inconsistent Image Sizes	Low	Resizing all images to a standardized size (e.g., 150x150 pixels) during preprocessing.
Kaggle Dataset	Potential for Incorrect Labels	High	Employ a manual review process for a subset of images to verify label accuracy. If errors are found, correct them or remove the affected images. Consider inter-annotator agreement to improve accuracy.

Overall Data Quality Assessment:

The initial assessment suggests a moderate level of data quality. Addressing the potential class imbalances and ensuring label accuracy are crucial steps before model training. The lower-severity issues related to image size and format will be resolved during the preprocessing stage.

A further assessment might be needed post-preprocessing to check the effectiveness of the implemented resolutions and to gauge the impact of data augmentation on class imbalance.