Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	SWTID1720033149
Project Title	Visual Diagnostics: Detecting Tomato Plant Diseases With Leaf Image Analysis
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarise 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 Dataset	Variations in Image Quality: Different resolutions and aspect ratios.	Moderate	Standardise image size by resizing all images to a consistent resolution, e.g., 256x256 pixels.
Kaggle Dataset	Background Clutter: Images may contain irrelevant background information that could distract the model from focusing on the key features of the tomato leaves.	Moderate	Image cropping removes irrelevant background and focuses on the regions containing the tomato leaves. This can be combined with a manual inspection to ensure that the regions of interest are correctly identified.

Kaggle Dataset	Missing or Incorrect Labels: Some images may be incorrectly labelled or lack labels.	Moderate	Perform a manual review and correction of the labels. Implement automated validation checks and cross-reference with expert annotations to ensure label accuracy.
Kaggle Dataset	Different Lighting Conditions: Images may be taken under varying lighting conditions, affecting the consistency of pixel values.	Low	Normalize pixel values to a specific range [0, 1] to reduce the impact of varying illumination conditions. Data augmentation techniques like brightness adjustment can also be applied to simulate different lighting conditions during training.
Kaggle Dataset	Non-representative Sample: The dataset may not cover all possible variations of leaf diseases.	High	Expand the dataset by sourcing additional images from different regions, seasons, and conditions. Implement domain adaptation techniques to improve model generalization.

Kaggle Dataset	Duplicate Images: Presence of duplicate images in the dataset.	Low	Use image hashing or similarity detection techniques to identify and remove duplicate images from the dataset.