Data Collection and Preprocessing Phase

Date	9 July 2024
Name	Pratik Prasad Mahadik
Project Title	Greenclassify: Deep Learning-Based Approach For Vegetable Image Classification
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

The project aims to develop and train a deep learning-based model for accurately classifying various types of vegetables based on their images. Data will be sourced from publicly available datasets on Kaggle, specifically a vegetable image dataset containing 21,000 images from 15 classes, each with 1,400 images in 224x224 resolution and in .jpg format. The dataset will be split into three parts: 70% for training, 15% for testing, and 15% for validation. The vegetable image dataset, available as a 560 MB zip file, can be accessed publicly via the provided Kaggle link.

Data Collection Plan Template

Section	Description
Project Overview	This project aims to develop and train a Deep Learning-Based model for vegetable image classification. The objective is to accurately classify various types of vegetables based on their images.
Data Collection Plan	Data will be collected from publicly available datasets related to vegetable images.(Kaggle)

Raw Data Sources Identified	In this dataset, there are 21000 images from 15 classes, where each class contains a total of 1400 images. Each class has an equal proportion and the image resolution is 224×224 and in *.jpg format.

	We split our dataset into three parts, where 70%(approx.) for training and 15%(approx.) for testing, and the rest 15%(approx.) for validation.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Vegetable Image Dataset	Dataset containing images of various types of vegetables for training and validation.	https://www.kaggle.com/datasets/misrakahmed/vegetableimage-dataset	Zip file	560 MB	Public