Data Collection and Preprocessing Phase

|  |  |
| --- | --- |
| Date | 21/06/25 |
| Team ID | SWTID1750180744 |
| Project Title | Smart Sorting: Transfer Learning for  Identifying Rotten Fruits and Vegetables |
| Maximum Marks | 2 Marks |

Data Collection Plan & Raw Data Sources Identification

Data Collection Plan Template

|  |  |
| --- | --- |
| **Section** | **Description** |
| Project Overview | This project aims to build an intelligent image classification system that can automatically detect whether fruits and vegetables are healthy or rotten using deep learning techniques. Leveraging the power of transfer learning with a pre-trained VGG16 model, the system is trained to recognize and classify images of produce into healthy or spoiled categories. The model is designed to assist in automated quality control, especially in agricultural sorting, retail supply chains, and food processing industries.   * Develop an image-based classification model using transfer learning (VGG16) to distinguish between healthy and rotten fruits/vegetables. * Preprocess and augment the dataset to improve model robustness and prevent overfitting. * Split the dataset into training, validation, and test sets to train and evaluate model performance accurately. * Evaluate the model using metrics like accuracy, loss, and test performance to validate its effectiveness. * Optimize the model by freezing base layers and training |

|  |  |
| --- | --- |
|  | only custom top layers for efficient fine-tuning. |
| Data Collection Plan | Kaggle https://[www.kaggle.com/datasets/muhammad0subhan/fruit-and-](http://www.kaggle.com/datasets/muhammad0subhan/fruit-and-) vegetable-disease-healthy-vs-rotten? select=Fruit+And+Vegetable+Diseases+Dataset |
| Raw Data Sources Identified | The Fruit and Vegetable Diseases Dataset consists of 28 directories, each representing a combination of healthy and rotten images for 14 different types of fruits and vegetables. This dataset is intended for training and evaluating machine learning models for disease detection. Images are categorized and organized to facilitate easy access and utilization for image classification and deep learning tasks.  Images in .jpg format   * Categories like Apple Healthy, Banana Rotten, Potato Healthy, etc. * Resolutions vary; mostly clear, high-quality images   Purpose in Project:  Used to train, validate, and test a CNN-based model for classifying produce based on its health condition. |

Raw Data Sources Template

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source Name** | **Description** | **Location/U RL** | **Format** | **Size** | **Access**  **Permission s** |
| Dataset 1 | This dataset has been compiled from various reputable sources, including Kaggle and  GitHub repositories. | https://[www.](http://www/) kaggle.com/ datasets/muh ammad0subh  an/fruit-and- | Image | Varies,after  rescaling it is (244,244) | Public |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Each image has been manually inspected and categorized to ensure high quality and relevance for the task of disease detection in  fruits and vegetables. | vegetable- disease- healthy-vs- rotten |  |  |  |