Data Collection and Preprocessing Phase

Date	11 July 2024
Team ID	SWTID1720067113
Project Title	Dog Breed Identification using Transfer Learning
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	The Dog Breed Identification project aims to classify images of dogs into their respective breeds using advanced machine learning techniques, specifically transfer learning. The primary objective is to leverage pre-trained convolutional neural networks to extract features from dog images and use these features to train a classifier that can accurately identify the breed of a dog from a given image.
Data Collection Plan	Data for this project will be sourced primarily from publicly available datasets. The main dataset will be obtained from the Kaggle competition "Dog Breed Identification," which provides a large collection of labeled dog images. This dataset is ideal for transfer learning as it contains a diverse set of images across various dog breeds.
Raw Data Sources Identified	**Description**: This dataset consists of images of dogs labelled with their corresponding breeds. The dataset includes a training set with labelled images and a test set with unlabelled images. The training

	set will be used to train the model, while the test set will be used to evaluate its performance. **Contents**: • train/: A folder containing training images. • test/: A folder containing test images. • labels.csv: A CSV file containing image IDs and their corresponding breed labels.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	Images of various dog breeds with labels	https://www.kaggle.com/competitions/dog-breed-identification/data?select=train	1. JPEG Images files 2. Labels are in CSV	1.1 GB	Public