Data Collection and Preprocessing Phase

Date	10 July 2024
Team ID	SWTID1720158677
Project Title	SportSpecs: Unraveling Athletic Prowess With Advanced Transfer Learning For Sports.
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 machine learning project aims to classify sports based on images using transfer learning. The objective is to build a high-accuracy model capable of real-time classification for 100 sports classes: cricket, wrestling, tennis, badminton, soccer, swimming, etc. This will be deployed as a web application using Flask.
Data Collection Plan	● Search for datasets related to sports activities, focusing on images of cricket, wrestling, tennis, badminton, soccer, swimming, and karate. ● Prioritize datasets with diverse and high-quality images to ensure robust model training. Augment the dataset with techniques like rotation, flipping, and scaling to increase diversity.

Raw Data Sources Identified	The raw data sources for this project include datasets obtained from platforms like Kaggle and UCI, which are popular for data science competitions and repositories. The provided sample dataset represents a subset of the collected information, encompassing images of various sports activities.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Dataset 1	Collection of sports images covering 100 different sports.. Images are 224,224,3 jpg format. Data is separated into train, test and valid directories. Additionallly a csv file is included for those that wish to use it to create there own train, test and validation datasets.	https://www.kaggle.com/datasets/gpiosenka/sports-classification	CSV	445 MB	Public
Dataset 2	The dataset is split into a training set and a test set. The training set consists of labeled images belonging to the following sports classes: cricket, wrestling, tennis, badminton, soccer, swimming, and karate.	https://www.kaggle.com/datasets/sidharkal/sports-image-classification/data	Image	877 MB	Private (with access)
		…	…	…	…