

1 DATASET COLLECTION

IMAGE CLASSIFICATION DATASET

Dataset: RSCD Road Surface Classification Dataset

Source: <https://ieee-dataport.org/documents/rscd-road-surface-classification-dataset-detailed-annotations-driving-assistance>

Number of Samples: The dataset contains 10,000 images.

Number of Class Labels: There are 8 class labels for road surface classification.

2 TEXT DATASET

Dataset: Pseudo-Label Generation for Multi-Label Text Classification

Source: <https://catalog.data.gov/dataset/pseudo-label-generation-for-multi-label-text-classification>

Number of Samples: The dataset includes a total of 1,000,000 samples.

Number of Class Labels: The dataset is designed for multi-label text classification tasks, and it contains 40 class labels.

3 TIME SERIES DATASET

Dataset: Weather Forecasting Time-Series Dataset

Source: <https://ieee-dataport.org/documents/weather-forecasting-time-series-dataset#files>

Number of Samples: The dataset consists of time series data collected hourly from 2014 to 2021, covering various weather attributes across multiple locations. The exact number of samples can be determined by the total count of data points across all locations and timestamps.

Number of Class Labels: This dataset is used for forecasting tasks, which are typically regression tasks rather than classification. Therefore, there are no class labels; instead, it contains continuous variables related to weather attributes (e.g., temperature, humidity, wind speed) that are predicted over time.

4 MEDICAL IMAGE DATASET

Dataset: OpenBHB: a Multi-Site Brain MRI Dataset for Age Prediction and Debiasing

Name: P.Satya Amrutha Raja

Roll Number: S20220010170

Source: <https://ieee-dataport.org/open-access/openbhb-multi-site-brain-mri-dataset-age-prediction-and-debiasing>

Number of Samples: The dataset contains 7,026 MRI scans.

Number of Class Labels: The dataset does not have predefined class labels for classification. Instead, it is used for age prediction, which is a regression task. Therefore, it does not have specific class labels but rather uses age as a continuous variable for prediction.

5 DNA SEQUENCE DATASET

Dataset: My Complete Genome

Source: <https://www.kaggle.com/datasets/zusmani/mygenome>

Number of samples: The dataset contains 100,000 genomic samples.

Number of class labels: There are 4 class labels in the dataset: 0, 1, 2, and 3.