A Supplementary Material

A.1 Data Cards

Table 2 shows the extracted documentation parameters from Kaggle and HuggingFace, which we categorized according to Datasheets [40].

On **HuggingFace**, we find information about the annotation creators (*e.g.*, crowdsource, experts, ml-generated) or specific task categories (*e.g.*, image-classification, image-to-text, text-to-image). Such parameters can be used to filter results when searching on HuggingFace, potentially enabling systematic analysis of a specific task or tag.

On **Kaggle**, we notice that some important parameters shown in the dataset website such as *temporal and geospatial coverage*, *data collection methodology*, *provenance*, *DOI citation*, and *update frequency* cannot be automatically extracted with their API, so we manually included them.

Kaggle automatically computes a *usability score*, which is associated with the tag "well-documented", and used for ranking results when searching for a dataset. Kaggle's *usability score* is based on:

- Completeness: *subtitle*, *tag*, *description*, *cover image*.
- Credibility: *provenance*, *public noteboook*, *update frequency*.
- Compatibility: *license*, *file format*, *file description*, *column description*.

The *usability score* is based on only 4 out of 7 aspects from Datasheets [40].

	Kaggle	HuggingFace
Motivation	*username* *dataset name* *description*	*username* *dataset name* *description*
Composition	*temporal coverage* *geospatial coverage*	*size categories*: $n < 1K$, $1K < n < 10k$, $1M < n < 10M$ *language*: en, es, hi, ar, ja, zh, ... *dataset info*: {image, class_label: bird, cat, deer, frog, ...} *data splits*: training, validation *region* *version*
Collection	*data collection method* *provenance*	*source dataset*: wikipedia, ... *annotation creators*: crowdsourced, found, expert-generated, machine-generated, ...
Preprocessing cleaning / labeling		
Uses		*task_categories*: image-classification, image-to-text, question-answering *task_ids*: multi-class-image-classification, extractive-qa, ...
Distribution	*license*: cc, gpl, open data commons, ... *DOI citation*	*license*: apache-2.0, mit, openrail, cc, ...
Maintenance	*update frequency*: weekly, never, not specified, ...	
Other	*keywords* *number of views* *number of downloads* *number of votes* *usability rating*	*tags* *number of likes* *number of downloads in the last month* *arXiv*

Table 2: Documentation parameters extracted from Kaggle and HuggingFace categorized according to Datasheets [40], except the last rows (Other). We represent in italic the extracted parameter, and show examples values for them. We include *description* in Motivation, although we find that this parameter can contain any type of dataset information.

A.2 Duplicates on Kaggle

We automatically retrieve all the duplicates for the top-10 listed MI datasets on Kaggle, as well as some popular datasets (suggested by the reviewers). In Table 3, we show the number of duplicates on Kaggle, the size of the original dataset, the cumulative size of the duplicates, and information about the license and description on Kaggle for the duplicates. We query the name of each dataset as shown in Table 3, except for DRIVE and NIH-CXR14. For NIH-CXR14, we use "nih chest x-ray" as query. When querying "DRIVE" (not case-sensitive) we got over 1800 datasets related to cars, Formula One, and similar topics. To refine results, we applied a case-sensitive filter, retaining only those with capitalized "DRIVE". We also queried Kaggle using "drive retina" and found 13 datasets, of which only 5 were new when compared to our filtered query. Combining the two set of results, we identified 41 duplicates.

Dataset	Duplicates	Size		License		Description
		Original	Kaggle	(%)	types	(%)
CheXpert [52]	47	440.0 GB*	342.1 GB	19.1	4	10.6
DRIVE [110]	34	30.1 MB	**11.7 GB**	26.5	5	85.3
fastMRI [59]	8	6.3 TB	215.2 GB	62.5	3	25.0
LIDC-IDRI [8]	43†	69.0 GB	**539.7 GB**	20.9	6	18.6
NIH-CXR14 [118]	47	42.0 GB	**654.6 GB**	59.6	5	97.9
HAM10000 [113]	141	3.0 GB	**468.4 GB**	42.6	11	26.9
MIMIC-CXR [58]	13	554.2 GB	62.1 GB	46.2	4	23.1
Kvasir-SEG [56]	51	66.9 MB	**8.7 GB**	41.2	4	15.7
STARE [49]	10	504.4 MB	**11.9 GB**	30.0	2	40.0
LUNA [105]	46	66.7 GB	**585.6 GB**	19.6	3	10.9
BraTS [80]	383	51.5 GB§	**7.3 TB**	30.0	9	92.4
ACDC [13]	28	2.3 GB	**127.7 GB**	28.6	5	14.3
ADNI [53]	70	N/A¶	803.3 GB	57.1	4	40.0
OASIS [61, 66, 78, 79]	53	34.5 GB‡	**657.7 GB**	56.6	3	15.1

Table 3: Information of the medical imaging dataset duplicates on Kaggle: number of duplicates; size of the original dataset and the storage on Kaggle; license information of the duplicates, percentage reported and different types of licenses; percentage of descriptions from duplicates that contain any text. *CheXpert dataset is 440 GB, however the 11 GB subset is the most commonly used and reshared. †We do not count LUNA duplicates for LIDC-IDRI. §BraTS datasets originated from challenges (2012-2022). These datasets are hosted at different websites and we couldn't retrieve their total size, dataset size is estimated from BraTS 2023. ¶The size details of the ADNI dataset were not readily available. We submitted an "ADNI Use Application" request but did not receive access in time. ‡OASIS dataset have 4 series, however, we only had access to the size information of OASIS-1 and OASIS-2, so the size estimation is based on these two series. We highlight in boldface when the cumulative size on Kaggle is larger than the original size. Data was collected in October, 2024.

We review each list and eliminate duplicates that are not relevant due to ambiguity, such as music datasets for OASIS. Some datasets were difficult to disambiguate because they contained no descriptions and provided compressed information (*e.g.*, npy files). We also found pretrained models listed under the dataset category. We decided to keep the examples we could not disambiguate and the pretrained models, as they were only a few. We keep duplicates that are aggregation of datasets, e.g. one instance groups together 3 different datasets for Alzheimer's, Parkinson's and "normal", which can cause data leakage [100]. LUNA was a challenge dataset created after LIDC-IDRI. We do not count LUNA-16 duplicates as duplicates of LIDC-IDRI, we only consider them for LUNA.