NEURAL INFORMATION PROCESSING SYSTEMS

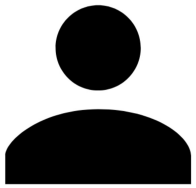
Copycats: the many lives of a publicly available medical imaging dataset

Amelia Jiménez-Sánchez
amji@itu.dk

Natalia-Rozalia Avlona

Dovile Juodelyte

Théo Sourget



Caroline Vang-Larsen

Anna Rogers

Hubert Dariusz Zając

Veronika Cheplygina

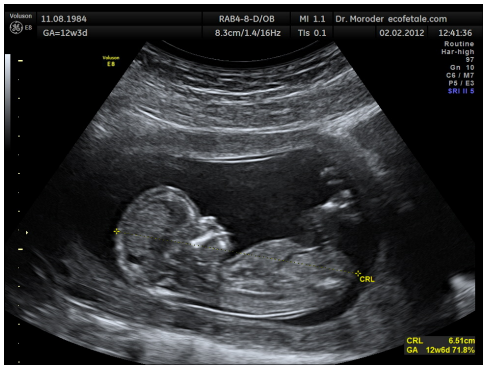
IT UNIVERSITY OF COPENHAGEN

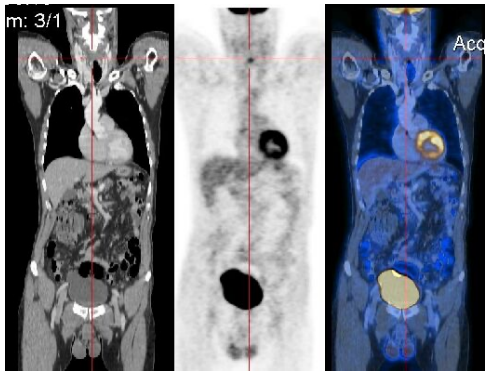
UNIVERSITY OF COPENHAGEN

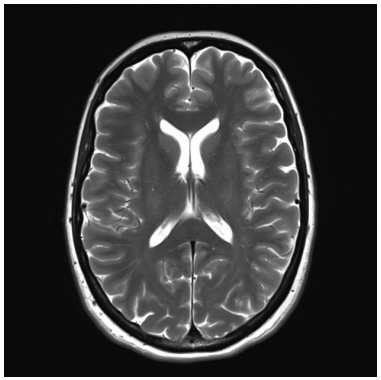
Datasets are fundamental

in ML and CV for understanding how algorithm performance impacts individuals, groups, or society.

MI datasets are crucial for safely realizing AI in healthcare.



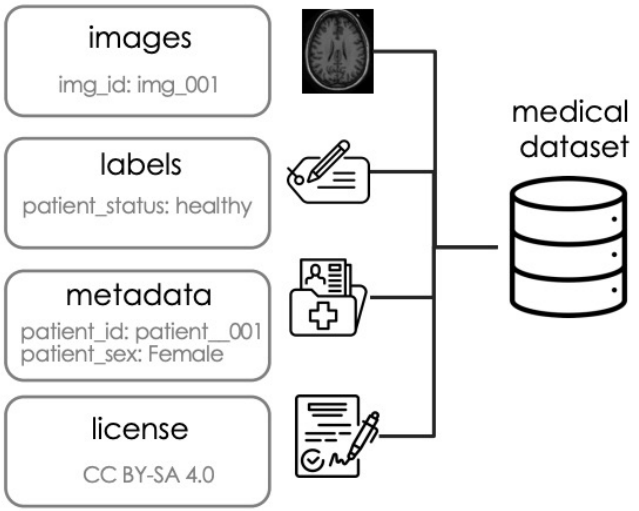




Source: Wikipedia

MI ≅ to CV, but:

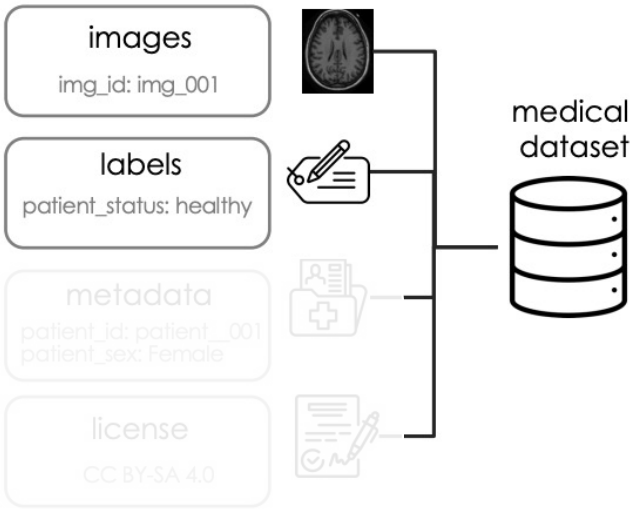
- de-identification for patient data,
- images from different patients,
- metadata: patient demographics or hospital scanner.



MI ≅ to CV, but:

- de-identification for patient data,
- images from different patients,
- metadata: patient demographics or hospital scanner.

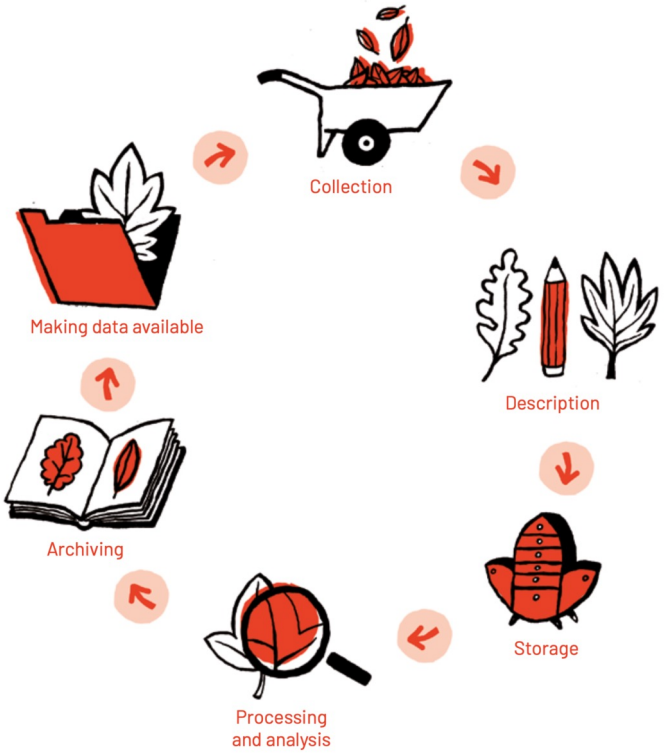
Treating MI as “small computer vision”: images + labels, while ignoring metadata can lead to unfair or inaccurate results.



Proprietary datasets --> public

Need for alternative models of **data governance**, **sharing** and **documentation**.

Community Contributed Platforms like Kaggle or HuggingFace enable public sharing of MI datasets… but



Source: Wikimedia – Data Management

Data management practices

Challenges:

- FAIR (Findable, Accessible, Interoperable, Reusable)[1]
- Tracking dataset versions and citations is difficult[2,3]
- Comprehensive documentation of dataset's lifecycle[4,5]

[1] Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3.1 (2016): 1-9.
[2] Peng, Kenny, et al. "Mitigating dataset harms requires stewardship: Lessons from 1000 papers." NeurIPS Datasets and Benchmarks Track (2021)
[3] Sourget, Théo, et al. "[Citation needed] Data usage and citation practices in medical imaging conferences." *MIDL* (2024).
[4] Gebru, Timnit, et al. "Datasheets for datasets." *Communications of the ACM* 64.12 (2021): 86-92.
[5] Hutchinson, Ben, et al. "Towards accountability for machine learning datasets: Practices from software engineering and infrastructure." *FAccT* (2021).

AMJI@ITU.DK

Study setup

We investigate dataset sharing, documentation and hosting practices for the 30 most cited CV, NLP and MI datasets.

Papers with Code	"Images"	"Text"	"Medical"
Modality	CV	NLP	MI

+ dataset distribution in:
  Community Contributed Platforms (CCPs): Kaggle, HuggingFace

1. Lack of persistent identifiers and storage

AMJI@ITU.DK

	Dataset	Original hosting source
1	CIFAR-10 [62]	cs.toronto.edu/kriz/cifar.html
2	ImageNet [95]	image-net.org
3	CIFAR-100 [62]	cs.toronto.edu/kriz/cifar.html
4	MNIST [68]	yann.lecun.com/exdb/mnist/
5	SVHN [78]	ufldl.stanford.edu/housenumbers/
6	CelebA [71]	mmlab.ie.cuhk.edu.hk/projects/CelebA.html
7	Fashion-MNIST [120]	github.com/zalandoresearch/fashion-mnist
8	CUB-200-2011 [110]	vision.caltech.edu/datasets/cub_200_2011/
9	Places [126]	places.csail.mit.edu
10	STL-10 [21]	cs.stanford.edu/ acoates/stl10/
1	GLUE [111]	gluebenchmark.com/
2	SST [102]	nlp.stanford.edu/sentiment/
3	SquAD [91]	rajpurkar.github.io/SQuAD-explorer/
4	MultiNLI [118]	cims.nyu.edu/sbowman/multinli/
5	iMDB reviews [74]	ai.stanford.edu/amaas/data/sentiment/
6	VQA [7]	visualqa.org/
7	SNLI [16]	nlp.stanford.edu/projects/snli/
8	Visual Genome [61]	homes.cs.washington.edu/[...]visualgenome
9	QNLI	gluebenchmark.com/ - derived from SQUAD
10	Natural Questions [63]	ai.google.com/research/NaturalQuestions
1	CheXpert [53]	stanfordmlgroup.github.io/competitions/chexpert/
2	DRIVE [104]	drive.grand-challenge.org
3	fastMRI [59]	fastmri.med.nyu.edu
4	LIDC-IDRI [8]	wiki.cancerimagingarchive.net/[...]pageId=[...]
5	NIH-CXR14 [112]	nihcc.app.box.com/v/ChestXray-NIHCC
6	HAM10000 [107]	dataverse.harvard.edu/[...]persistentId=doi[...]
7	MIMIC-CXR [58]	physionet.org/content/mimic-cxr/2.0.0/
8	Kvasir-SEG [56]	datasets.simula.no/kvasir-seg/
9	STARE [50]	cecas.clemson.edu/ ahoover/stare/
10	LUNA [99]	luna16.grand-challenge.org

1. Lack of persistent identifiers and storage

AMJI@ITU.DK

	Dataset	Original hosting source
1	CIFAR-10 [62]	cs.toronto.edu/kriz/cifar.html
2	ImageNet [95]	image-net.org
3	CIFAR-100 [62]	cs.toronto.edu/kriz/cifar.html
4	MNIST [68]	yann.lecun.com/exdb/mnist/
5	SVHN [78]	ufldl.stanford.edu/housenumbers/
6	CelebA [71]	mmlab.ie.cuhk.edu.hk/projects/CelebA.html
7	Fashion-MNIST [120]	github.com/zalandoresearch/fashion-mnist
8	CUB-200-2011 [110]	vision.caltech.edu/datasets/cub_200_2011/
9	Places [126]	places.csail.mit.edu
10	STL-10 [21]	cs.stanford.edu/ acoates/stl10/
1	GLUE [111]	gluebenchmark.com/
2	SST [102]	nlp.stanford.edu/sentiment/
3	SquAD [91]	rajpurkar.github.io/SQuAD-explorer/
4	MultiNLI [118]	cims.nyu.edu/sbowman/multinli/
5	iMDB reviews [74]	ai.stanford.edu/amaas/data/sentiment/
6	VQA [7]	visualqa.org/
7	SNLI [16]	nlp.stanford.edu/projects/snli/
8	Visual Genome [61]	homes.cs.washington.edu/[...]visualgenome
9	QNLI	gluebenchmark.com/ - derived from SQUAD
10	Natural Questions [63]	ai.google.com/research/NaturalQuestions
1	CheXpert [53]	stanfordmlgroup.github.io/competitions/chexpert/
2	DRIVE [104]	drive.grand-challenge.org
3	fastMRI [59]	fastmri.med.nyu.edu
4	LIDC-IDRI [8]	wiki.cancerimagingarchive.net/[...]pageId=[...]
5	NIH-CXR14 [112]	nihcc.app.box.com/v/ChestXray-NIHCC
6	HAM10000 [107]	dataverse.harvard.edu/[...]persistentId=doi[...]
7	MIMIC-CXR [58]	physionet.org/content/mimic-cxr/2.0.0/
8	Kvasir-SEG [56]	datasets.simula.no/kvasir-seg/
9	STARE [50]	cecas.clemson.edu/ ahoover/stare/
10	LUNA [99]	luna16.grand-challenge.org

Without DOI, access to (meta)data is uncertain, which is problematic for reproducibility.

2. Vague licenses

AMJI@ITU.DK

	Dataset	Distribution terms (use, access, sharing)	License	Please cite this paper
1	CIFAR-10 [62]		□ Unspecified	Yes
2	ImageNet [95]	Terms of access	□ Unspecified	Yes
3	CIFAR-100 [62]		□ Unspecified	Yes
4	MNIST [68]		□ Unspecified	Yes
5	SVHN [78]		□ Unspecified	Yes
6	CelebA [71]	Agreement	□ Unspecified	Yes
7	Fashion-MNIST [120]		● MIT	Yes
8	CUB-200-2011 [110]		□ Unspecified	Yes
9	Places [126]		△ (C), ● CC-BY	Yes
10	STL-10 [21]		□ Unspecified	Yes
1	GLUE [111]		See original datasets	Yes
2	SST [102]		□ Unspecified	Yes
3	SquAD [91]		● CC-BY-SA 4.0	No
4	MultiNLI [118]		● Various CC	Yes
5	iMDB reviews [74]		□ Unspecified	Yes
6	VQA [7]	Terms of use	△ (C), ● CC-BY	Yes
7	SNLI [16]		● CC-BY-SA 4.0	Yes
8	Visual Genome [61]		● CC-BY 4.0	Yes
9	QNLI		□ Unspecified	No
10	Natural Questions [63]		● CC-SA 3.0	No
1	CheXpert [53]	Research Use	□ Unspecified	Yes
2	DRIVE [104]		□ Unspecified	No
3	fastMRI [59]	Sharing Agreement	□ Unspecified	Yes
4	LIDC-IDRI [8]	TCIA Data Usage	● CC-BY-3.0	Yes
5	NIH-CXR14 [112]		□ Unspecified	Yes
6	HAM10000 [107]	Use Agreem.	● CC-BY-NC-4.0	Yes
7	MIMIC-CXR [58]	Phys. Use Ag. 1.5.0	● PhysioNet 1.5.0	Yes
8	Kvasir-SEG [56]	Terms of use	□ Unspecified	Yes
9	STARE [50]		□ Unspecified	Yes
10	LUNA [99]		● CC-BY-4.0-DEED	Yes

No clear license or terms of use, but … “Please cite this paper”.

They are often missing when distributed within CCPs.

2. Vague licenses

AMJI@ITU.DK

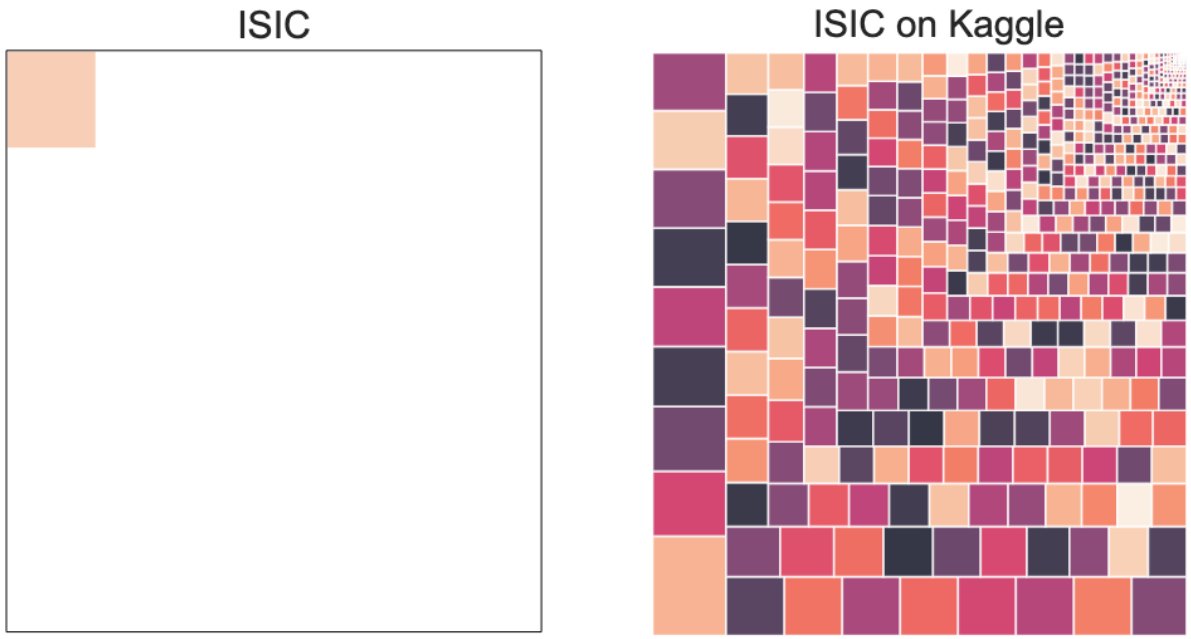
	Dataset	Distribution terms (use, access, sharing)	License	Please cite this paper
1	CIFAR-10 [62]		□ Unspecified	Yes
2	ImageNet [95]	Terms of access	□ Unspecified	Yes
3	CIFAR-100 [62]		□ Unspecified	Yes
4	MNIST [68]		□ Unspecified	Yes
5	SVHN [78]		□ Unspecified	Yes
6	CelebA [71]	Agreement	□ Unspecified	Yes
7	Fashion-MNIST [120]		● MIT	Yes
8	CUB-200-2011 [110]		□ Unspecified	Yes
9	Places [126]		△ (C), ● CC-BY	Yes
10	STL-10 [21]		□ Unspecified	Yes
1	GLUE [111]		See original datasets	Yes
2	SST [102]		□ Unspecified	Yes
3	SquAD [91]		● CC-BY-SA 4.0	No
4	MultiNLI [118]		● Various CC	Yes
5	iMDB reviews [74]		□ Unspecified	Yes
6	VQA [7]	Terms of use	△ (C), ● CC-BY	Yes
7	SNLI [16]		● CC-BY-SA 4.0	Yes
8	Visual Genome [61]		● CC-BY 4.0	Yes
9	QNLI		□ Unspecified	No
10	Natural Questions [63]		● CC-SA 3.0	No
1	CheXpert [53]	Research Use	□ Unspecified	Yes
2	DRIVE [104]		□ Unspecified	No
3	fastMRI [59]	Sharing Agreement	□ Unspecified	Yes
4	LIDC-IDRI [8]	TCIA Data Usage	● CC-BY-3.0	Yes
5	NIH-CXR14 [112]		□ Unspecified	Yes
6	HAM10000 [107]	Use Agreem.	● CC-BY-NC-4.0	Yes
7	MIMIC-CXR [58]	Phys. Use Ag. 1.5.0	● PhysioNet 1.5.0	Yes
8	Kvasir-SEG [56]	Terms of use	□ Unspecified	Yes
9	STARE [50]		□ Unspecified	Yes
10	LUNA [99]		● CC-BY-4.0-DEED	Yes

No clear license or terms of use, but … “Please cite this paper”.

They are often missing when distributed within CCPs.

3. Duplicates datasets and missing metadata

AMJI@ITU.DK



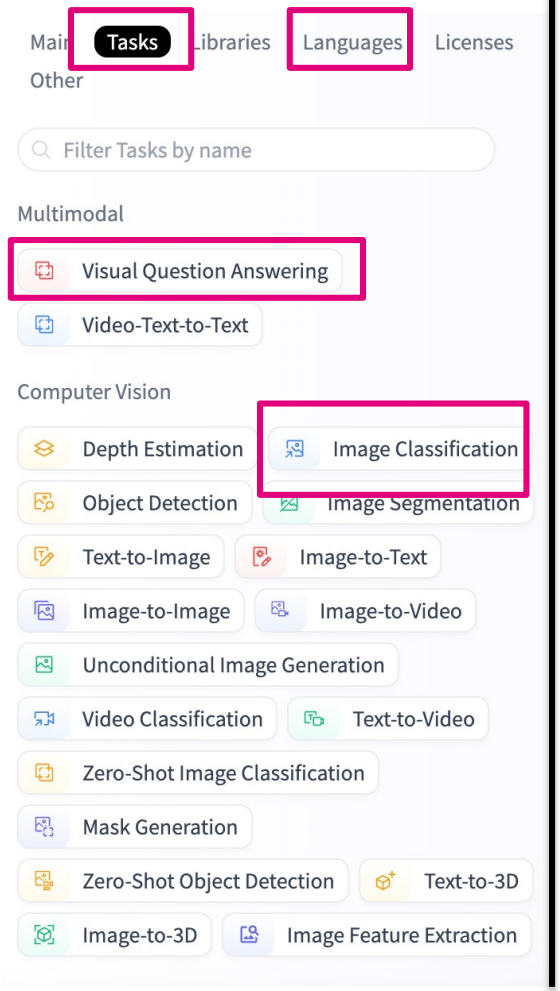
38 GB

640 datasets, **2.35 TB!!!** – May 2024

4. Where are the datasheets?

AMJI@ITU.DK

HuggingFace



Kaggle

About Dataset

Summary description

- The PAD-UFES-20 dataset was collected along with the Dermatological and Surgical Assistance Program (in Portuguese: Programa de Assistência Dermatológica e Cirurgica - PAD) at the Federal University of Espírito Santo (UFES-Brazil), which is a nonprofit program that provides free skin lesion treatment, in particular, to low-income people who cannot afford private treatment.
- The dataset consists of 2,298 samples of six different types of skin lesions. Each sample consists of a clinical image and up to 22 clinical features including the patient's age, skin lesion location, Fitzpatrick skin type, and skin lesion diameter.
- The skin lesions are: Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), Bowen's disease (BOD), Melanoma (MEL), and Nevus (NEV). As the Bowen's disease is considered SCC in situ, we clustered them together, which results in six skin lesions in the dataset, three skin cancers (BCC, MEL, and SCC) and three skin disease (ACK, NEV, and SEK)
- All BCC, SCC, and MEL are biopsy-proven. The remaining ones may have clinical diagnosis according to a consensus of a group of dermatologists. In total, approximately 58% of the samples in this dataset are biopsy-proven. This information is described in the metadata.

Usability
6.47

License
Other (specified in description)

Update frequency
Unspecified

Tags
Image | Classification | Medicine

HuggingFace provides documentation that is more thorough and well-organized than Kaggle's.

4. Where are the datasheets?

AMJI@ITU.DK

HuggingFace

Kaggle

About Dataset

Summary description

- The PAD-UFES-20 dataset was collected along with the Dermatological and Surgical Assistance Program (in Portuguese: Programa de Assistência Dermatológica e Cirurgica - PAD) at the Federal University of Espírito Santo (UFES-Brazil), which is a nonprofit program that provides free skin lesion treatment, in particular, to low-income people who cannot afford private treatment.
- The dataset consists of 2,298 samples of six different types of skin lesions. Each sample consists of a clinical image and up to 22 clinical features including the patient's age, skin lesion location, Fitzpatrick skin type, and skin lesion diameter.
- The skin lesions are: Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), Bowen's disease (BOD), Melanoma (MEL), and Nevus (NEV). As the Bowen's disease is considered SCC in situ, we clustered them together, which results in six skin lesions in the dataset, three skin cancers (BCC, MEL, and SCC) and three skin disease (ACK, NEV, and SEK)
- All BCC, SCC, and MEL are biopsy-proven. The remaining ones may have clinical diagnosis according to a consensus of a group of dermatologists. In total, approximately 58% of the samples in this dataset are biopsy-proven. This information is described in the metadata.

Usability ⓘ
6.47

License
Other (specified in description)

Update frequency
Unspecified

Tags

Image | Classification | Medicine

Kaggle computes a *usability score*, related to the "well-documented" tag.

4. Where are the datasheets?

AMJI@ITU.DK

HuggingFace

Main Tasks Libraries Languages Licenses Other

Filter Tasks by name

Multimodal

Visual Question Answering

Video-Text-to-Text

Computer Vision

Depth Estimation Image Classification

Object Detection Image Segmentation

Text-to-Image Image-to-Text

Image-to-Image Image-to-Video

Unconditional Image Generation

Video Classification Text-to-Video

Zero-Shot Image Classification

Mask Generation

Zero-Shot Object Detection Text-to-3D

Image-to-3D Image Feature Extraction

Kaggle

About Dataset

Summary description

- The PAD-UFES-20 dataset was collected along with the Dermatological and Surgical Ass
Portuguese: Programa de Assistência Dermatológica e Cirurgica - PAD) at the Federal Unive
Brazil), which is a nonprofit program that provides free skin lesion treatment, in particular, to
cannot afford private treatment.
- The dataset consists of 2,298 samples of six different types of skin lesions. Each sample
and up to 22 clinical features including the patient's age, skin lesion location, Fitzpatrick ski
diameter.
- The skin lesions are: Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Acti
Seborrheic Keratosis (SEK), Bowen's disease (BOD), Melanoma (MEL), and Nevus (NEV). As
considered SCC in situ, we clustered them together, which results in six skin lesions in the d
(BCC, MEL, and SCC) and three skin disease (ACK, NEV, and SEK)
- All BCC, SCC, and MEL are biopsy-proven. The remaining ones may have clinical diagnos
of a group of dermatologists. In total, approximately 58% of the samples in this dataset are
information is described in the metadata.

Usability

This score is calculated by Kaggle.

Completeness · 50%

- ✗ Subtitle
- ✓ Tag
- ✓ Description
- ✗ Cover Image

Credibility · 67%

- ✓ Source/Provenance
- ✗ Public Notebook
- ✓ Update Frequency

Compatibility · 50%

- ✓ License
- ✓ File Format
- ✗ File Description
- ✗ Column Description

ified in description)

equency

Classification

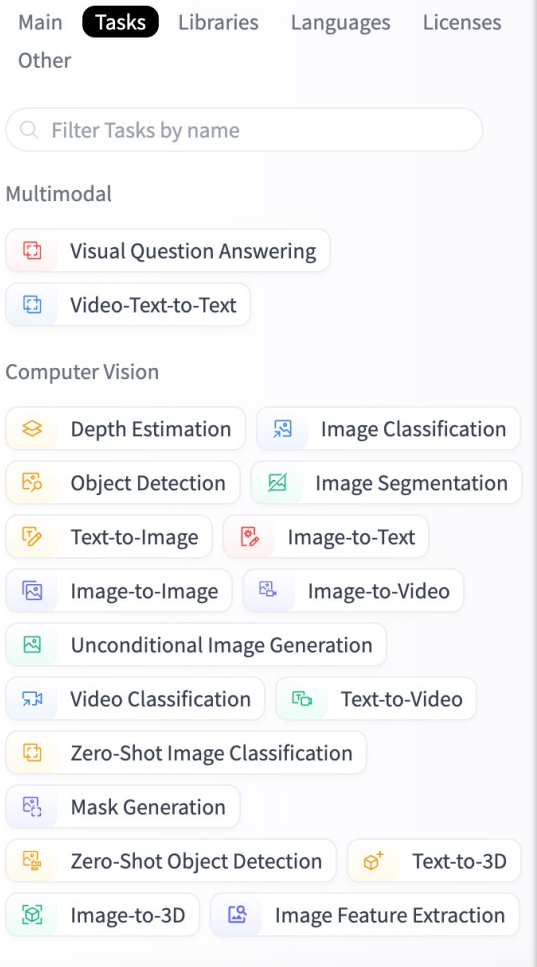
Users can obtain a high score w/o key information:

- *update frequency*: "never"
- *provenance: "uses internet sources"*

4. Where are the datasheets?

AMJI@ITU.DK

HuggingFace



Kaggle

About Dataset

Summary description

- The PAD-UFES-20 dataset was collected along with the Dermatological and Surgical Assistance Program (in Portuguese: Programa de Assistência Dermatológica e Cirurgica - PAD) at the Federal University of Espírito Santo (UFES-Brazil), which is a nonprofit program that provides free skin lesion treatment, in particular, to low-income people who cannot afford private treatment.
- The dataset consists of 2,298 samples of six different types of skin lesions. Each sample consists of a clinical image and up to 22 clinical features including the patient's age, skin lesion location, Fitzpatrick skin type, and skin lesion diameter.
- The skin lesions are: Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), Bowen's disease (BOD), Melanoma (MEL), and Nevus (NEV). As the Bowen's disease is considered SCC in situ, we clustered them together, which results in six skin lesions in the dataset, three skin cancers (BCC, MEL, and SCC) and three skin disease (ACK, NEV, and SEK)
- All BCC, SCC, and MEL are biopsy-proven. The remaining ones may have clinical diagnosis according to a consensus of a group of dermatologists. In total, approximately 58% of the samples in this dataset are biopsy-proven. This information is described in the metadata.

Usability
6.47

License
Other (specified in description)

Update frequency
Unspecified

Tags
Image | Classification | Medicine

Efforts to integrate data documentation 🥐, many fields are left empty by the users 😿.

AMJI@ITU.DK

Recommendations

🔒 Access to datasets: predictable, open licensing, and persistent.

👀 Evaluation: including rich metadata & real-word evaluations.

📖 Documentation: complete and up-to-date.

🤝 CCPs could benefit from commons-based governance.

TL; DR: Promote **better data governance** in the context of **MI datasets** to mitigate risks and uphold the **reliability** and **fairness** of AI models in **healthcare**.