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

Amelia Jiménez-Sánchez, Natalia-Rozalia Avlona, Dovile Juodelyte, Théo Sourget, Caroline Vang-Larsen, Anna Rogers, Hubert Dariusz Zając, Veronika Cheplygina


✉️ amji@itu.dk, vech@itu.dk
🐈 https://purrlab.github.io/

Not just "small computer vision"!

Medical Imaging (MI) datasets are crucial for the **safe** implementation of **AI** in **healthcare**.

- **Open data** is important for progress in community.
- MI datasets have special properties: multiple images per patients, metadata (demographics, hospital scanner,...).
- Shared datasets on community-contributed platforms (CCPs) like Kaggle or HuggingFace (HF) often ignore this information.
- Missing such metadata can lead to **overoptimistic performances**, and **adverse outcomes** for patients.

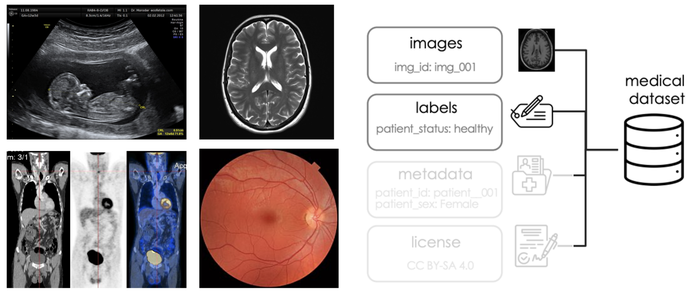

Figure 1: Treating MI as general CV while ignoring metadata regarding patient splits or hospital scanners may lead to unfair or inaccurate results.

Sharing practices, bad for reproducibility

We query Papers with Code and select the top-10 datasets for **CV**, **NLP**, and **MI**.

	Finding	Issue
Hosting	CV, NLP: author/university websites	not FAIR*
	MI: grand-challenges, PhysioNet	
	w/o persistent id & storage	uncertain access
Licenses	missing for most CV	author attribution
	$\approx 50\%$ for NLP, MI	

**FAIR: Findable, Accessible, Interoperable, Reusable.*

Duplicate datasets and missing metadata

- Duplicates with **no documentation** or source citation.
- Group together 3 datasets for Alzheimer's, Parkinson's and "normal", which can cause **data leakage**.
- **Breast cancer**: INBreast dataset (x24).
  "I'm just uploading here this data as a backup"
- **Skin lesions**: ISIC (**x640!**); PAD-UFES-20 (x10), includes one instance containing ISIC data.

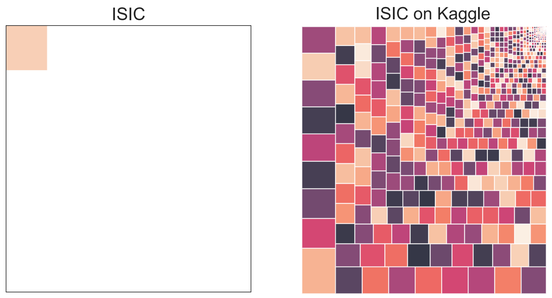

Figure 2: Representation of the storage size for ISIC (skin lesion) data: 38 GB original (left) vs 640 versions with 2.35 TB on Kaggle (right).

Where are the datasheets?

Composition and collection are the most represented fields, while motivation, preprocessing, and usage are often missing.

	Pros	Cons
CCPs	metadata format 🥐	users leave the fields empty
Kaggle	*usability score*: doc	↑ score w/o information
	update frequency	"never"
	provenance	"uses internet sources"
HF	*task_categories*: uses,	
	systematic analysis	

Recommendations

- 🔒 **Access**: predictable, open licensing, and persistent.
- 👀 **Evaluation**: including rich metadata and emphasizing real-world evaluations to reveal biases or shortcuts.
- 📖 **Documentation**: complete and up-to-date.
- 🤝 CCPs could gain from **commons-based governance**, with roles like *data administrator*, and *data steward*.

Acknowledgement - DFF Inge Lehmann 1134-00017B.

UNIVERSITY OF COPENHAGEN

IT UNIVERSITY OF COPENHAGEN