Preliminary Project Statement

Investigating Duplicate Medical Imaging Datasets on Kaggle

Popular dataset platforms such as Kaggle and Hugging Face offer a wide variety of datasets for research and development. A common challenge, however, is the commonness of duplicate datasets. These duplicates often lack comprehensive data descriptions or contain only partial data, which can lead to issues in research validity and reproducibility.

This project aims to investigate duplicate medical imaging datasets in Kaggle, focusing initially on the **The International Skin Imaging Collaboration (ISIC)** [1] [2] with room for expanding on different platform such as Hugging Face if times promises. The **ISIC** dataset, notably, includes hundreds of duplicates, highlighting the need for this research.

Inspiration was drawn from the research by my supervisors' Veronika Cheplygina's research group, where they started exploring this topic already in Jiménez-Sánchez et al. [3] and from the research by Cassidy et al. [4].

In this project, I aim to identify a list of methods and variables for detecting duplicates and compare these duplicates to their original datasets. The plan is to try to use several approaches, including **image size comparison**, where the dimensions and file size of images are compared, **pixel-by-pixel comparison**, which checks for exact matches in pixel values, and **hashing techniques**, which involve generating hashes of the images to quickly detect exact or near-exact duplicates. The outcomes of this project include a comprehensive list of methods for detecting duplicates, an understanding of the differences between these duplicate datasets, and an analysis of what kind of overlap exists.

Additionally, if time allows, I plan to conduct a baseline model tests on two nearly identical datasets, one original and one duplicate, to compare their outcomes and perceive any differences in results.

References

[1] International Skin Imaging Collaboration, "ISIC Challenge Datasets." International Skin Imaging Collaboration, 2020. doi: 10.34970/2020-DS01.

[2] "ISIC | International Skin Imaging Collaboration," ISIC. Accessed: Sep. 12, 2024. [Online]. Available: https://www.isic-archive.com

[3] A. Jiménez-Sánchez *et al.*, “Copycats: the many lives of a publicly available medical imaging dataset,” Jun. 10, 2024, *arXiv*: arXiv:2402.06353. doi: 10.48550/arXiv.2402.06353.

[4] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, “Analysis of the ISIC image datasets: Usage, benchmarks and recommendations,” *Med. Image Anal.*, vol. 75, p. 102305, Jan. 2022, doi: 10.1016/j.media.2021.102305.