**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,** [**https://www.isic-archive.com/**](https://www.isic-archive.com/)**,** [**https://challenge.isic-archive.com/data/**](https://challenge.isic-archive.com/data/)**) )** 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 (Jiménez-Sánchez et al. <https://arxiv.org/abs/2402.06353v2>) and from the research by Cassidy et al. <https://www.sciencedirect.com/science/article/pii/S1361841521003509>.

In this project, I aim to identify a list of some methods and variables for detecting duplicates and to compare these duplicates to their original dataset. The outcomes of this project include a comprehensive list of methods for detecting duplicates, understanding the differences between this 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.