PAD-UFES-Group I dataset report

Project Group I members:

Jamie Underwood, David Malinaitis, Anastassia Zinovjev,
Samrat Rijal, and Madalena Iordachi

February 18, 2026

The PAD-UFES-20 dataset (hereon simply referred to as 'the dataset') is a collection of over two thousand images collected via phone camera of skin lesions. These clinical images were gathered for the express purpose of providing varied public clinical data on various skin conditions as comparable data is either private (such as the 'skin vision' dataset) or is dependent on complex and expensive tools. Exceptions to this are the HAM10000 and BCN20000 datasets.

The dataset was compiled in part by the Federal University of Espírito Santo's Dermatological and Surgical Assistance Program from patients within the Espírito Santo region of Brazil. This region is highly populated by European immigrants with a large contingent of patients identifying as being of Pomeranian descent, a region between Germany and Poland. Additionally, Espírito Santo is a highly rural region in which the majority of patients are or were farmers. As such, it is worth noting that the dataset is non generalized outside of elderly farmers of European descent who were exposed to extended sunlight during their work.

Observations of our section of the dataset show group I's data to be relatively clear. Photos are not blurry, lesions are clearly visible and are at most only partially covered by hair or ink markings on the skin. Lighting is inconsistent but not unmanageably so. These observations are all in line with the publication of the dataset as low quality samples, samples with identifying information, samples with obscured lesions were all removed prior to publication. The samples are all raw images without modification as the intent was to provide training data that would match with what computer aided diagnosis systems will be likely to encounter under real life conditions, hence the lighting.

The metadata provided along with the dataset is extensive covering 21 attributes such as family history of cancers, age, sex, exposure to common carcinogens such as smoking, drinking and pesticide use, region of descent for both mother and father, access to clean water and human waste disposal, and data about the patients skin type along with notes about the lesion(s) and if there is a supporting biopsy. It is worth noting for some samples this data is missing or unknown with only age, biopsy status and region being guaranteed to be present along with cataloging factors such as patient id.

Our link to the github is:
https://github.com/JamieUnderwoo/2026-PDS-GroupI/tree/main